Computational Modelling of Transitive Inference:

A Microanalysis of a Simple Form of Reasoning

M.R. Harris

Ph.D.

University of Edinburgh

1988

I DECLARE THAT THIS THESIS HAS BEEN COMPOSED BY MYSELF AND THAT THE WORK DESCRIBED IN IT IS MY OWN:

(M.R. Harris)

This thesis is dedicated
to my father, Alan Harris.

Acknowledgements

My supervisors, Alan Bundy and Brendan McGonigle, have guided and criticised this work. I appreciate their continued enthusiasm for interdisciplinary research, as well as for my selected project. Dr Margaret Chalmers has also supported the collaboration, and has been very helpful in sorting out relevant data, analyses and presentation. Dr. Francis Provan advised on the analysis of variance of the reaction time data in chapter 5. Thanks to all those who gave time to read or discuss my work, including Alan Smaill, Brian Cartwright, Paul Brna, MaryAngela Papalaskari, Steve Owen, Lincoln Wallen, Robert Inder, John Jones and, in particular, Clive Wynne.

Abstract

The goal of this research is to provide an account of transitive inference which has both psychological and computational justification, and to relate it to the broader context of inference and information gaining systems in general. Perhaps the most well known transitive inference task is called the 'N-term series problem' which has been used, for example, to assess cognitive development in young children. It is argued that this task taps basic cognitive skills which are likely to form the building blocks of more complex forms of reasoning.

In a typical five-term series task, subjects are given the information $A > B$, $B > C$, $C > D$ and $D > E$, where the letters denote arbitrary stimuli and '$>$' denotes an ordinal comparison, such as 'longer than'. Subjects are then able to infer the relationship between 'remote' pairs such as B and D. A typical phenomenon associated with this, and related tasks, is called the ordinal distance effect — the time taken to make comparisons between remote pairs is, typically, faster than the responses to the original training pairs, suggesting that much inference has taken place during the initial learning process.

Recent evidence from monkeys has been shown to be fully representative of this class of experiment. Furthermore, the monkey studies are the only ones to provide a sufficiently rich database to permit a microanalysis based on computational modelling. This thesis contains such an analysis, and it is shown how many aspects of the subjects' behaviour can be accounted for with a surprisingly simple rule-based model, in which subjects' strategies are represented by highly constrained rule stacks. The model can account for the major phenomena associated with the five-term series task, and can model individual subject variation within a principled framework. Finally, an algorithm is proposed for acquiring appropriate rule stacks, given a random sequence of training examples such as received by experimental subjects.

Table of Contents

List of Figures

List of Tables

Chapter 1

Motivation and Overview

Chapter one explains the motivation for and the contents of this thesis. It can be viewed as being a guide, with each section corresponding to one of the subsequent chapters. Section numbers in this chapter thus correspond to chapter numbers in the rest of the thesis. Also, citations are avoided where the relevant chapter supplies the detail.

1.1 Motivation and Scientific Context

In attempting to construct intelligent software for machines, it makes sense to try and model biological systems, because these are the only 'proven' working examples of intelligent, adaptable behaving systems we have. Our current understanding is that nervous systems not only serve transducing functions but also act as information processing and storage mechanisms. As information processors, higher organisms (particularly humans) seem so flexible that it is not clear what kinds of constraints might be imposed by their physical characteristics. Indeed, it seems likely that the physical organisation of the brain is subservient (over the course of evolution) to the dictates of information processing needs.

The question is, what are these dictates? Some of them might be expressed in engineering terms such as efficiency and robustness, parsimony, modularity *etc.* These are constraints on processing *algorithms* once their function has been defined. The second source of constraint must be in terms of *function*, as governed by the role of the organism in the environment and the information available.

Our goal is therefore interdisciplinary — to work towards discovering the principles that govern information gain and the control of action in behaving systems.

A key idea behind this research is that there exist certain 'primitive' or 'precursor' cognitive skills which cross age and species boundaries and which form the basis on which more complex skills and strategies are built.

1.1.1 Existing Approaches

The obvious places to look for insights into natural systems would appear to be Psychology and the Brain Sciences. However, these have not always been motivated towards reporting characteristics which can be implemented. Such investigations have tended to produce coarse abstractions of group behaviour, verbal 'introspective' reports or causal models using the 'brain metaphor'. By the latter, I mean theories which use discovered or hypothetical physical characteristics of the brain to provide a metaphor for information processing. For example, stimulus-response theory and its derivatives originated from physical analogies. Another example is the 'spreading activation' theories of memory. Even where such accounts are valid, by appealing to a particular physical architecture, researchers have tended to overlook the epistemological reasons why information might be organised in a particular way. After all, as has already been suggested, the architecture most likely evolved to deal with particular kinds of information processing problems, and not *vice versa.*

On the other side of the coin, we have the 'pure' information processing approach of Logic, Computer Science and 'basic' (mainstream) (Bundy, 1986) Artificial Intelligence (AI). This type of research might be described as formally specifying information processing problems and constructing algorithms to solve them. The weak link here is not in the process of constructing algorithms; AI has a powerful and expanding set of techniques for solving well formalised problems. The difficulty is in asking the right questions in the first place, as people's intuitions are often misleading not only about how they solve problems but also about *what* problems they solve.

Working artificial systems still tend to be 'domain limited' and the problem solving and learning techniques embodied in these programs are (where they have been abstracted) rather weak methods. Indeed, the weakness of the abstracted techniques has led some to hypothesise that 'domain knowledge' is all important and that the main business of AI is to find ways of representing 'world knowledge'. Whilst this is undoubtably a useful enterprise, and has led to the fields of Knowledge Engineering and the attempts to formalise 'common sense reasoning', both the source and the test of the representations are the researchers' own intuitions about what they know. Similar problems exist with problem solving and other aspects of 'basic' AI; the field is really an engineering discipline aimed at developing techniques and formal methods. Whilst there is nothing wrong with this, there is no guarantee that this alone will achieve our goal of understanding and emulating biological systems.

To summarise the overall research problem, it is not the case either that the AI researcher can receive straightforward guidance as to what to implement or that the role of the psychologist is to simply choose between the algorithms provided by AI and decide which is most appropriate for describing his/her data. Rather, there must be a symbiotic relationship between the two disciplines. Unfortunately such inter-disciplinary research programmes are relatively rare. The fields do import concepts from one another but there is relatively little collaboration at the early investigative stages.

On their own, both AI and Psychology have failed to understand the engineering principles underlying intelligent, adaptable behaving systems. This may well simply be due to the fact that the class of solutions to the problems of information gain and management is small, and thus not easy to arrive at. This would suggest that biological systems, which have evidently solved the problems, are likely to share common fundamental principles, whether by inheritance or by convergent evolution. This justifies our interest in comparative psychology. There are many striking examples of common biological solutions, ranging from digestive to optical systems. Our first step is to identify a domain which we

suspect will help us gain insight into the less tangible problem of information management.

1.1.2 Inference

One aspect of the observed rationality of behaving systems is their ability to make inferences, and there has been much interest over the years in formalising various forms of inference. In common usage, to 'infer' means to 'go beyond the information given' and this word, or its equivalent in other languages, was certainly used long before the invention of mathematical logic. A huge variety of mental activities can be classified as 'inference', ranging from the intellectual to the subconscious. For discussion purposes, I have found it useful to distinguish between (approximately) three[1] levels of inference and these are described below, although the distinctions do blur into one another. Also, some inferences can become faster with practice, perhaps even changing category, although there are limits to this.

- Conscious, or *problem solving* kinds of inference are perhaps the best understood. Firstly, they are open to introspection; people can, for example, write books about how to calculate what cards your poker opponents are likely to be holding. Secondly, they have been subject to extensive automation, especially in recent years with the advance of computer science and artificial intelligence (AI). Thirdly, they have been investigated by cognitive psychologists of the 'state space search' school, as described in section 1.2. This kind of inference is also sometimes referred to as *planning* and is the slowest of the three, with time-courses of a few minutes to many hours.

- Inference which is much faster (lasting a few seconds) is not so easily decomposed or isolated, and the investigator must resort to creating artificial

[1] A fourth level, involving low-level perceptual and motor processing (automatic processes) is outside the scope of this paper.

tasks and using group data or multiple measures to aid building a theory about the underlying processes. These will be referred to as *fundamental* forms of inference (for it seems likely that their basic components form the building blocks for more complex inferences) and last from a fraction of a second to a few seconds. Examples of this are linguistic inference, such as anaphoric resolution, single digit arithmetic, categorical (*eg* reasoning about set inclusion) inference and ordinal comparisons.

- *Composite* forms of inference come midway between the two forms above, and it is debatable whether they should comprise a separate category. However, there do appear to be some kinds of inference which are multi-step and yet there is only partial (conscious) access to the intermediate results and the nature of the representations used in solution. Examples of this are multi-digit arithmetic, understanding spatial descriptions, constructing a representation of a series, and syllogistic inference. These kinds of inferences take a few seconds to a few minutes to complete. They have also been investigated by cognitive psychologists but the state-space approach is less appropriate and more indirect methods have to be used.

How are the faster types of inference to be explained? Many have been led to conclude that there is some kind of innate mental logic which mediates such forms of inference (Inhelder & Piaget, 1958). Johnson-Laird devotes a significant proportion of his book *Mental Models* (Johnson-Laird, 1983) to arguing against the 'doctrine of mental logic'. One of his arguments, is that postulating a logic engine in the head is simply inadequate as an explanation. Firstly, it does nothing to explain learning processes and, secondly *control* of inference is left unexplained. A logic engine will only draw *correct* inferences, but it is unlikely to draw *useful* inferences unless directed to do so by some controlling element. Examination of the cognitive psychology and AI literature leads one to the conclusion that there is a definite lack of adequate study and (computational) explanation of, in particular, fundamental forms of inference.

Deduction and Induction

Another, traditional, classification of inference is into deductive and inductive forms. In the broadest sense, deduction is inferring specifics from generalities and induction[2] is the converse process, going from specific instances to generalisations. Most types of reasoning involve both processes at some stage. For example, think of a rat which has learned to solve some class of maze problems (finding its way to a food source). In deciding whether to turn left or right at any particular junction, it can be said to be deducing which way to turn from the knowledge it has acquired about the maze, the task and from local information about the junction. On the other hand the process of acquiring such generalisations must be one of induction, learning from the successes and failures of previous decisions. Presumably, unless the rat derived a perfectly satisfactory algorithm for finding its way through the maze, the learning process would continue indefinitely, and in parallel with the deductive processes.

It seems unfortunate, then that most psychological studies of inference have concentrated (at least superficially) on deductive inference alone, a point made by (McGonigle & Chalmers, 1986). However, as the authors also point out, in the process of trying to isolate a deductive act, many studies have uncovered interesting phenomena associated with the integration of the information necessary to perform the tasks. Thus light is thrown on the more general mechanisms of information management, including induction. Chapter 7 deals with some of the issues related to deduction in more detail.

1.1.3 The Contribution of AI

Although the traditional Artificial Intelligence (AI) methods of using intuition and introspection to uncover the nature of inference are of limited use (worse, they can be misleading), AI and computer science do provide a body of work-

---

[2]Induction in the common and philosophical senses, not mathematical induction.

ing *behaving systems* which can form a rich source for analogies and models, together with an understanding of their formal properties. On the other hand, new techniques in psychology involve a *microanalysis* of performance with repeated measures of more than one kind. This is explained more fully later on, the point for the moment being that it is possible, as in Physics, to obtain multiple measures on an unseen object and to build up a picture of it by indirect methods. The difference between the kinds of models sought here and the mathematical models of Physics is that a cognitive model should be similar *in kind* to its subject material, which we assume is also a computational process. Our challenge, then, is to construct process models of fundamental forms of inference informed by and (in turn) informing a multiple level empirical analysis. Psychology contributes by providing techniques for catching and observing the inference occurring *in vivo* and AI can offer new ways of formalising theories of process and representation.

1.1.4 Transitive Inference

Transitive inference is, in general, a form of reasoning about rankings. Formally, it can be specified by an inference rule: 'if x is more than y and y is more than z then x is more than z'. The word 'more' has been used here to stand as proxy for any kind of comparison along a scale. In logic, this is referred to as transitivity of inequality as distinct from symmetrical or 'associative transitivity'. This latter type of inference involves reasoning about equivalence and it will be mentioned again later in the context of related psychological research. Transitive inference can occur without applying an inference rule of the kind above. In general, we can say that it occurs wherever the 'given' information relates pairs of items and the 'inferred' information relates novel combinations of items *as if* they were part of an overall order. Transitive inference is a suitable candidate for investigation (ideally as part of a coordinated programme of research into fundamental forms of inference) for a number of reasons:

1. It is fast. Ordinal comparisons can be made in one or two seconds and this is too fast for introspection into the sub-processes involved.

2. It is abstract. Transitive inference is one of the simplest forms of reasoning about *relations* between symbols. The inference is thus one level removed from the perception or mental representation of the objects themselves. This means that our findings will not be domain limited and that we can reasonably hope that the concepts evolved will be extensible to other forms of inference.

3. It has an important place in theoretical psychology. Transitive reasoning was initially studied in the context of child development. Piaget (Piaget & Inhelder, 1969), for example, was interested in the emergence of 'formal' reasoning. There is also a group of related kinds of inference which have been studied to various extents, for example, seriation, comparison of items in long-term memory, learning of temporal sequences and linguistic comprehension of comparatives.

4. Empirical data is available. This is related to the last point in that there are a number of well documented phenomena that have emerged from studies in transitive inference. More importantly for this study, however, there is a large body of data on monkeys' performance of the transitivity task which has been made available for this investigation (see section/chapter 3), the use of which will be justified further on.

5. It is ubiquitous. Transitive inference is likely to arise in any system which reasons about items which differ from each other with respect to some dimension. A metrical representation (associating a quantity with each element to be compared) may be inappropriate or impractical if information about the items is relative or uncertain and so an ordinal representation (ranking) is likely to be used. Examples are reasoning about preferences, relative positions of objects in space or time, social orderings, size orderings etc. Of these perhaps *preference orders* are the most general and could be

regarded as more 'primitive' than the others. For example, even the simplest organisms discriminate between situations which are better or worse for obtaining food.

6. Controlling transitive inference is an important problem in AI. Transitive inference is ubiquitous in AI for the same reasons as given above. The obvious examples involve reasoning about physical systems (size, length, mass *etc*). There are also 'meta-level' inferences which need to be made, for example, heuristic search involves imposing a preference order onto the problem space, planning involves reasoning about sequences of actions, etc. Use of a transitive inference rule in an unconstrained way can lead to very large search spaces being generated and so anything that can be learned about the control of transitive inference is likely to be useful.

7. Although it is traditionally thought of as a deductive form of inference, the task can also be thought of as involving the integration of information. This means we can potentially learn something about the principles of information management employed by subjects.

It should be pointed out at this point that there are really two levels to transitive inference as it occurs in the typical transitive inference task. The first is a slow *problem solving* or *composite* type of inference during which the subject learns about a series and the second is the ordinal comparison, which is fast. Discussion of the slower aspect, which can be regarded as subsuming the second (and is thus a harder problem), will be deferred until chapter 7 (section 1.7).

1.2 Background: Methods and Models

Meanwhile, chapter 2 describes some previous psychological and computational approaches to inference, with the emphasis on relating the *kinds* of model obtained to the *kinds* of data and methods used. This provides a context for our own methodology. In particular, studies of potential relevance to transitive inference are reviewed.

1.2.1 The Transitive Inference Paradigm

A typical three term series problem, as used by Piaget in developmental studies of children, is shown below:

> 'Edith is fairer than Suzanne and Edith is darker than Lilly. Who is the fairest, Edith, Suzanne or Lilly?'

Some early models of how children perform such reasoning were based on the idea of language-like categories. For example, 'Edith is fairer than Suzanne' could be naively represented by slotting each person into one of the two categories suggested by the comparative. Thus Edith is *fair* and Lilly is *dark*. If the latter half of the conjunction is similarly represented, then only one name is unambiguously categorised as *fair* (Lilly). Suzanne is labeled as dark and Edith is neither dark nor light. More sophisticated models were proposed for adults which, nevertheless, also worked on the the same basic principle of decomposing comparatives into 'primitives'.

At the same time, another school of thought was that such problems were solved by employing mental imagery. It was argued that the comparatives were used to order terms into a 'mental line' which could be inspected (in the mind's eye) for the purpose of drawing conclusions. Thus, in the previous example, the line 'Suzanne-Edith-Lilly' might be constructed. Obviously there is also the need to label the line with a direction, in this case, dark to light.

			$A>C$
	$A>B$		$B>D*$
	$B>C$		$C>E$
Given:	$C>D$	Infer:	$A>D$
	$D>E$		$B>E$
			$A>E$

Table 1–1: A 5-term series problem

Although the 'imagist' type models are less parsimonious, they gained ground with the generalisation of the three-term problem to N terms. A five-term series problem is illustrated schematically in table 1–1. A series containing five or more terms contains at least one pair containing items which cannot be uniquely categorised. In the example, this is the pair marked with an asterix. If the continuum is size, for example, neither B nor D can be uniquely categorised as 'big' or 'small'. Studies of five term, and longer, versions of the N-term series problem are therefore crucial.

Bryant and Trabasso developed a version of the N-Term series task which involved the subject comparing the relative lengths of rods referred to only by colour. One of the main motivations was to show that children could perform transitive inference at a much earlier stage than previously thought if they were given appropriate non-linguistic stimuli.

However, Trabasso went on to become interested in the *process* by which knowledge of a series was acquired and by the nature of the representation of the series itself. To this end he used two important techniques. The first was to monitor subjects' performances when they had only partially learned the series and the second was to take *multiple measures* of responses in the same task using both *reaction time* and error rate indices. A major phenomenon to emerge was the *ordinal distance effect*, which is that the speed of comparison of two items in a series is, on average, proportional to their separation. For example, judging

the relative position of adjacent items is *slower* than comparing items which are separated by intervening elements of the series.

Trabasso *et al* thought that (what we call) the ordinal distance effect was a version of an analogous kind of distance effect which is found in a different paradigm. The latter is called the *symbolic distance effect* (SDE), and there are grounds for thinking that it should be treated separately. The 'comparisons' involved in the SDE were, in the first studies, size comparisons of familiar objects referred to only by name. The effect obtained was that the discriminations were faster the bigger the difference in size between the (imaginary) objects. The SDE has since been found to be ubiquitous in memorial comparisons along virtually any kind of perceivable or imaginable dimension.

In fact, the SDE is not a single phenomenon but a compound of related effects. It is much more complex than the ordinal distance effect, involving linguistic and world knowledge, rather than abstract representations constructed at the time of the task. Due to this complexity, and motivational and methodological differences between researchers, there is a lack of a solid corpus of data suitable for microanalysis. Whilst the possibility is not precluded that similar mechanisms underly problem solving in the N-term series problem and the SDE paradigm, this thesis concentrates on the former.

Bryant and Trabasso's 'non-linguistic' version of the N-term series task opened the door for a new *representation* oriented approach to transitive inference which clearly separated linguistic from other reasoning abilities. If very young children can reason about 'order' information, combining separately presented chunks of information in a logical fashion, this would appear to indicate that linguistic competence does not necessarily form the basis of abstract reasoning abilities. The task is clearly tapping some basic mechanism for organising information, and our goal is to find out what this mechanism is.

Pairs	
No reward	Rewarded
yellow	blue
blue	white
white	green
green	red

Objects are always presented in these pairs with the left-right position of the rewarded item varied. Subjects are then tested to see if they spontaneously choose between novel pairs according to the implied series 'yellow, blue, white, green, red' (actual colours vary between subjects).

Table 1–2: Example training scheme for five-term series.

1.3 Non-verbal Transitive Inference

Five and six series experiments have now been carried out under a wide variety of conditions, with different subjects, species and materials. It has now been established beyond reasonable doubt that subjects can perform crucial 'internal' comparisons (BD in the previous example) and that the ordinal distance effect is a robust phenomenon. Chapter 3 reviews some of this work, including experiments with children, monkeys and, recently, pigeons. The fact that analogous behaviour arises in different species reinforces the idea that there are limited solutions to the information management problems involved. However, there is only one study which has produced detailed enough data to enable the micro-analysis of individual subjects, and this provides the main focus of this chapter, and for the remainder of this thesis.

McGonigle and Chalmers adapted Bryant and Trabasso's task for use with monkeys. This was done for a number of reasons, some of them already mentioned, but the main point here is that, non-human subjects could be tested more intensively and over a longer period, allowing high density data to be collected, with the potential for follow-up studies on the same subjects. Children, on the other hand, are more difficult to motivate in such a repetitious task.

It was found that monkeys could perform this task, and produced a performance profile remarkably similar to that of children. Table 1–2 shows the basic training scheme. Subjects were repeatedly presented with the training pairs until they could reliably pick the rewarded colour and were then tested. As well as testing the subjects on novel pairs from the series (*eg* blue and green in the example), 'triadic' tests were introduced, which involved the subjects choosing from a three element subset of the series.

Chapter 3 describes the obtained phenomena in more detail, but the key features are that the spontaneous bias on all the pairs was highly transitive (after the adjacent pairs had been learned) and that every individual subject showed an ordinal distance effect. Previously, these effects had only been statistically significant for *groups* of subjects. The demonstration of the ordinal distance effect in individuals effectively proves that the monkey subjects do not make remote comparisons on the basis of stored training pairs.

Performance on the triadic tests was significantly worse, however. When subjects transferred to this second task they had a tendency to sometimes choose 'second best'. In other words, although each three element subset could, in principle, be totally ordered according to the implied series, in practice, subjects would often take the middle item (from the ordered subset).

For example, given the series A, B, C, D and E, where the 'correct' choice from any pair is the one near the E end, then the 'transitive' choice from the triad ABD is D. However, B would be chosen on many trials. If the task seems trivial using this notation, remember that the left-right order of the choice stimuli (colours) are scrambled. Finally, a surprising additional result was that subjects showed a spontaneous improvement on the triadic tests when they were given an extra intensive testing session. This was without any intervening retraining and without selective feedback on their responses.

Some of the features of this data which make it attractive as the basis for computational modelling are its regularity, the large quantity of data on each individual subject and the availability of both reaction time and choice (error) measures. From the theoretical perspective, the experiments also have an advan-

tage in that there are two basic tasks which were given to subjects after training, giving two windows onto the underlying processes.

1.4 The Stack Model

It was decided to use the monkey data to inform a modelling attempt of the N-term series task. Existing models, such as Trabasso's and Breslow's (discussed in chapter 6), concentrated on accounting for the ordinal distance effect, and made no predictions about triadic choice patterns. The mechanism described below [3] was a fresh start. As well as being an efficient way of representing a series, the mechanism appeared simple enough to be plausibly employed by monkeys and made some interesting predictions about triadic choices. In outline, it is based around the concept of *avoidance* and *selection* rules. Each rule is a conditional, with the the left hand side stating where the rule is relevant and the right side stating a 'choice tactic'. A strategy for performing the task consists of having a small stack of rules (four), each one of which is an instruction to either avoid or select a particular feature, in along with a 'control strategy' which tests the condition of each rule in turn, and applies the first relevant one it comes across.

For example, the series *Yellow*, *Blue*, *White*, *Green*, *Red* can be represented by the rules below. This strategy lends itself to a production system notation, and is explained formally in chapter 4.

1) If *Red* is present then select *Red.*
2) If *Yellow* is present then avoid *Yellow*
3) If *Green* is present then select *Green.*
4) If *Blue* is present then avoid *Blue.*

Briefly, this mechanism (henceforth referred to as the stack model) was selected as a candidate for further investigation for the following main reasons:

---

[3]First proposed in (Harris, 1985).

1. It is simple, and therefore ought to be tested on the Occam's razor principle, as well as being plausible for monkeys.

2. It can be extended to make predictions about the three choice situation in the triadic task. Choices to the 'middle' item can be explained by unsuccessful application of an *avoidance* rule which only provides an unambiguous choice in the binary situation. The assumption was made that where the choice specified by the rules was ambiguous, a random choice would be made.

3. A number of different combinations and orders of rules can successfully perform the task, thus there is some potential for accounting for variation between subjects.

Chapter 4 goes on to describe how the stack model satisfies (qualitatively) some of the constraints implied by the phenomena characteristic of the five-term series task, such as the *ordinal distance effect.*

1.5 Evaluating the Stack Model

Chapter 5 deals with the quantitative evaluation of the stack model. The approach is to assess the 'fit' at coarse levels of description and then to progressively chunk the data more finely until the inadequacies of the model become apparent. The analysis begins with triadic error patterns and goes on to look at reaction times during binary tests.

Given that the stack model is the only one to make clear predictions about triadic tests, the first step was to re-examine the monkey triadic data to see if the patterns of errors could be accounted for. A novel feature of the approach taken was to map out a space of possible triadic choice profiles based on the variations of rules allowed in the stack model. These projected profiles were then compared with summaries of the actual data obtained for *individual* subjects. The net result was that individual monkeys' choice patterns could be better

characterised by the operation of individual rule stacks than by previous 'group' types of models. It is interesting to compare the approach taken here with the 'state-space' methodology described in section 1.2. Both approaches rely on the idea of a behaviour space, but whereas the latter compares projected and obtained *sequences of actions*, we compare choice *profiles*, as a necessary consequence of the type of data we are dealing with.

The intensive additional triadic tests were analysed and it was found that the improvement could be modelled by a change in subjects' control strategies, whilst each individual maintained the same stack of rules.

Up to this point, no use had been made of the reaction time (RT) data except for the original acceptance criterion of the model that it should account for the ordinal distance effect. The stack model makes a clear prediction about RT patterns for individuals, which is that the time to respond should increase with the depth of the rule which determined the choice outcome. The binary RTs were therefore plotted accordingly, and it was found that, for five out of six subjects, there was a significant linear correlation between reaction time and depth of rule. For two of the subjects, nearly all of the variation in RTs could be accounted for in this fashion. This is called the *depth* effect.

The net conclusion from the evaluation is that the stack model is a good approximation of monkeys' decision procedures during the five-term series task, although one subject may be employing a deviant strategy. However, the assumption that subjects are using a rigidly ordered set of rules may be partially incorrect. Rule order may change slightly, over the course of time or it may have a stochastic component to it.

1.6 Deductive Processes in Transitive Inference

Having presented an account of how monkeys solve the N-term series problem, chapter 6 compares the stack model with the alternatives. The stack model is shown to be the most parsimonious model which adequately characterises the published phenomena. The relationship between models is pointed out by representing the different mechanisms in a common formalism.

It is argued that most rely, at some level, on hypothesising the existence of a transitive inference rule and deductive reasoning. In short, the stack model is the most parsimonious model which adequately characterises the published phenomena. As Trabasso's and Breslow's models are the main rivals to the stack model, some of the theoretical objections to them are summarised below.

Trabasso *et al*'s model of transitive inference was based around the idea that subjects explicitly constructed a series or associated objects with a pre-existing series. The retrieval and comparison process he proposed was analogous to visual discrimination (as studied in psychophysics) where items close together are more easily confused.

Breslow models Trabasso's data in a more parsimonious way and his account is computationally complete. The basic idea is that subjects construct two *associative chains*, each starting from one end of the series. Staying with our size example, the subject would start by identifying the objects that were unambiguously big or small (the end points) and chain from these items using the contiguity information in the training pairs. The chain used for information retrieval would depend on the direction of the comparative used in the question, and works by scanning the chain until one of the items to be compared is located.

Breslow had a number of misconceptions about the nature of his model, which he thought showed that subjects could perform the task in a 'non logical' manner. For example, he thought of the 'chains' as being purely associative but they are in fact clearly directional both in construction and retrieval and so the

subject is using more than purely the contiguity relationship from the training examples. Ironically, however, it turns out that Breslow's model can be captured by a set of logical rules, and it is shown that the model belongs to a family of similar representations.

Bundy proposed a model in which a series is represented by a discrimination tree. Although the 'default' form of this model is not compatible with the phenomena, it is shown that both Breslow's mechanism and the stack model can be thought of as employing tree-like representations.

Finally, many of the models conflate the N-term series and SDE phenomena; to some, extent, they try to model them both. The remainder of chapter 6 examines the models with respect to the *marking* and *congruity* effects. It is found that it is very difficult for simple representations to show both of these effects at the same time, without assumptions being made extrinsic to the computational needs of the decision process. A very simple mechanism is described which does show both of these effects, but it cannot simultaneously account for the phenomena of the N-term series task other than the gross distance effect. It is concluded that Trabasso may have been wrong in his assumption that the N-term series task tapped the same mechanisms as employed to make the pseudo-perceptual comparisons in the SDE effect. However, more research needs to be done before the connection (or lack of it) between the SDE and N-term series effects can be fully established.

1.7 Transitive Inference as Induction

Chapters 5 and 6 demonstrate that the stack model has descriptive power with respect to the monkey data. Although evaluation of this kind of performance model could be taken further, this should not be done at the expense of asking *why* subjects in the N-term series task make transitive choices at all, when they could simply store the four adjacent pairs? During testing, the monkey subjects were rewarded indiscriminately, so the reason must be entirely to do with repre-

sentation. Part of the answer may be that the stack model is, computationally, cheap enough to be a sensible representation of four pairwise relations. However, this still leaves the problem of how the representation is learned and why it varies from subject to subject.

We cannot assume either that the subjects have a ready made transitive inference rule or that, even if they had, that they would know to apply it in this context. Without such a rule being added to the axioms representing the task, the transitive relationships between remote pairs simply do not follow deductively from the original pairs. Even with a transitive inference rule, there still remain parts of the task which are essentially inductive in character. For example, subjects would have to infer what the finite set of training examples is.

Furthermore, the simple stack model can only go so far in modelling the subtleties of subjects' behaviour. A possible rationale for its limitations is that they are due to the restrictive framework imposed by the deductive inference paradigm. Working within such a framework, there is a tendency to try and justify any proposed mechanism solely on the grounds of parsimony and computational efficiency. When the observed mechanism appears to violate these constraints, we are at a loss as to what additional complexity to import. It was decided to try out the idea that the mechanism which learns and performs the transitive inference task derives from one which deals with a generic kind of multiple choice situation.

It was decided to approach these problems by asking what kind of basic cognitive tools would be needed to deal with the problem of learning how to identify and deal with an unclearly defined set of choice situations? The model which emerged uses components of the stack model, such as the concept of 'avoidance' and 'selection' as primitives but casts them in a broader role of choice tactics which could be employed in a wide variety of choice situations. Similarly, the rules change status to being considered as hypotheses about the set of choice situations faced by the subject.

A program has been written which implements some basic principles of this theory. It can be shown that this inductive learning algorithm appears to learn

about the series in an 'end-inwards' fashion as found by Trabasso in his subjects and by McGonigle and Chalmers with the monkeys. The stacks that are generated depends on the history of previous choices that the program has made, and this is used to explain why different subjects employ different stacks. This model has not been empirically tested against individuals' training data but it is offered as lending plausibility to the stack account and providing a possible framework for future work.

1.8 Conclusions

Chapter 8 draws together the conclusions from the modelling work and suggests further tests and possible extensions of the model. Finally, the significance of the modelling work is discussed with respect to wider issues in development and learning.

1. The motivation of this research is to identify and model basic cognitive skills which cross age and species boundaries and which could form the building blocks of more complex skills in biological or synthetic behaving systems.

2. Transitive inference, as found in the in the N-term series task, appears to be such a basic form of reasoning.

   (a) It is sub-linguistic and cross-species.

   (b) It is fast, and closed to introspection.

   (c) It is abstract and thus not domain limited.

   In general, transitive inference is also interesting from the point of view of synthesis, as it is ubiquitous in formal systems.

3. A model of (this kind of) transitive inference was developed based on a microanalysis of monkey data. The microanalysis breaks down into approx-

imately two stages, though the development of the model was inevitably partially an iterative process.

(a) A computational mechanism, the *stack model*, was proposed as candidate for modelling subjects' decision processes during the N-term series task. The mechanism is simple enough to be plausibly employed by animals and can be justified on the grounds of computational efficiency and parsimony.

(b) As well as qualitatively satisfying the criteria of the published N-term series phenomena, the model stood up to detailed quantitative evaluation against data from seven monkey subjects.

4. A mechanism which subsumes the stack model was proposed, by which subjects could learn to perform the N-term series task. This was implemented as a computer program, and found to have the right gross learning characteristics. Although this model has not been evaluated in detail, it lends further plausibility to the stack model account.

5. The role of transitive inference as a 'pre-logical' primitive is discussed, with respect to information management and human development.

Chapter 2

Background: Methods and Models

This chapter gives some of the background of previous attempts to model various kinds of inference. The purpose of this is to illustrate how psychological and AI approaches have previously been combined and to help to provide a context for our own methodology. The modelling attempts selected from the literature have largely been chosen for their relevance to the transitive inference paradigm or for their historical significance. The more psychological approaches are dealt with first, followed by those with a computational emphasis. A critical comparison of of models of transitive inference is postponed until chapter 6.

2.1 Characterising Inference

From a psychologist's point of view, the main motivation in studying inference is to *understand* how and why a behaving system can appropriately 'go beyond the information given'. From the AI perspective, the goal is to actually *create* behaving systems. Both disciplines face the problem of understanding, for example, the effects of incomplete information and the role of world knowledge. So we see that, at this level, the aims of AI and psychology are perfectly compatible; systematic observation of the behaving systems can reveal constraints in their information processing and creating information processing systems can give rise to guidance as to what observations to make.

If we assume, as was posited in chapter 1, that there exist general cognitive structures (which exist regardless of an individual's state of learning), and

that these structures can be described in computational terms, then two lines of investigation are needed. One is to try and discover things about the cognitive structures themselves, and the second is to explain how situation-specific information processing strategies are generated from these general structures. Perhaps inevitably, most psychological and AI studies of reasoning have tended to concentrate on discovering what strategies can be employed in specific situations. It is difficult to work backwards from such task specific descriptions of process to find the basic cognitive skills. Nevertheless, it is worth bearing in mind as a top level goal.

This brings us onto the question of the relationship between formal and behavioural characterisations of inference. Mathematical logic and computer science provide us with convenient formalisms for describing behaving systems and, in particular, decision making and inference systems. **There is no *a priori* prescription as to how this should be done or for how such systems should be compared with humans.** However, there are certain existing inference systems (both computational and logical) created to perform in particular task situations where they are used as a tool. With these systems there is precedent: many people have already drawn the analogy between the steps of logical inference that a logician makes and the corresponding steps (or series of sub-steps) in a formal proof. As this fit seems quite good, this leads onto the question of whether there is a one to one correspondence between all human inference and logical inference by rule? Do we need to characterise human inference any further or can we just look for the appropriate rules (*albeit* a difficult task) and let these act as our model?

Before considering these questions further, let us observe that the correspondence between a logician's chain of inference and the corresponding proof of a formal system is not particularly surprising since the formal system was created to perform the same task. Imagine any machine producing an act which could also be performed by humans: the mechanism of the machine could be posited as a model for the mechanism employed by the human. However, there will certainly be tasks which the human mechanism will find 'equivalent' or adapt-

able to, but which the machine mechanism will find impossible. For example, a food processor is fine for stirring soup but shows no transferral whatsoever to serving it. So it has a failing as a model of the human arm holding a spoon. Even a chess machine is similarly dedicated to an exact task specification. The inferences it makes are quite useless if the rules of the game are changed slightly, whereas humans typically show adaption and transfer. Similarly, a simple automatic theorem prover is quite likely to disappear into an endless search if it is given certain types of input, a behaviour which would cause the unfortunate demise of a human logician.

Measuring the degree of transferral between different tasks is a useful psychological probe for finding salient or relevant task features and thus establishing a 'task ecology' — a domain of related tasks. For example, in early experiments on the animal learning, it was discovered that rats could be trained to choose a grey square in preference to a black one, thus demonstrating that they can discriminate between different light intensities. But it is not immediately obvious what is being learned:

(i) Grey $\Longrightarrow$ Reward

(ii) darker(STIMULUS1, STIMULUS2) $\Longrightarrow$ Reward

In (i) a very simple association of a property with reward is learned but in (ii) a comparison between two stimuli is necessary. Which is learned can be tested by measuring transferral to a related task in which the animal has to choose between a grey and white square. Similarly, we can investigate what happens when the shape of the stimulus is changed etc. In this simple example, the possibilities for different mechanisms facilitating the same performance profile are somewhat restricted but with more complex tasks, and with more complex animals, it can be all too difficult to know just what has been learned. Also, in more complex cases, learning need not just be affected by the way the input is 'coded' but also by the strategy adopted by the subject as affected by the task in hand or anticipation of the kind of information it will receive. There is a danger of the

researcher assuming that, because the subject responds 'correctly', the subject has learned the very concept that the experimenter used to judge correctness.

To return to the previous theme, if we want computers to be able to make inferences in the same *way* as humans (because humans are good at it) then, even within a very restricted domain, we should not expect a program to behave bizzarely with new tasks which humans often cope with using, essentially (and subjectively), the same strategy as they started with.

For this to be possible, a new methodology is necessary. It is unlikely to be fruitful, without guidance, to generate and test mechanisms for possible psychological validity. Rather we must look to human behaviour itself in the first instance and discover, by subjective and empirical means (aided by theory and models), what types of behaviour appear to fit together in a descriptive taxonomy. For example, it is unlikely to be fruitful to posit the same mechanism for two behaviours which can clearly be distinguished on psychological grounds (for example, if there is an order of magnitude difference in their time-courses). Also, two superficially different behaviours may turn out to have analogous behavioural profiles.

Bearing these general issues in mind, let us look at some actual examples.

2.2 Johnson-Laird's Constructivist Account of Inference

In his ambitious work *Mental Models* (Johnson-Laird, 1983), Johnson-Laird offers a comprehensive framework for understanding inference. He falls approximately into the psycholinguistic category of Clark and Banks (see below), in that he was interested in inference and representation from the viewpoint of *comprehension* of text. His main theoretical paradigm was that of *constructive semantics* as opposed to *interpretive semantics*. The basic dichotomy can be characterised by asking the question whether a piece of text is 'understood' with

the use of *forwards* or *backwards* inference. In this context, a subject is said to understand a piece of text if he can answer questions about it and 'go beyond the information given', that is, make textual inferences. The question is, do we store the premises contained in the text in a 'superficial' (near linguistic) form and then perform deductions on them as required, or do we make deductions *as part of the comprehension process*, storing the results in some kind of integrated representation?

Johnson-Laird's major emphasis is on how the syntactic and logical form of textual premises could affect the form of mental representation. He used two main domains in his investigations. The first was syllogisms and the second was spatial descriptions. These fall approximately into the *composite* category oof inferences, as described in section 1.1.2. The basic experimental framework was the same for both domains: the subject was given a number of premises (sentences) and asked to perform a task afterwards, which could be to draw a 'spontaneous' conclusion, verify an inference or to remember the original text. In the spatial description domain, the initial 'verification' was done pictorially, but this does not alter the general framework. The type of data he collected was basically similar to that of Clark's paradigm, in that he varied the syntactic and logical form of the text and measured the effect on error rates (and sometimes the average reaction time to respond). He also used a 'confusion' measure in which subjects had to judge similarity of various mutated forms of a text to their memory of the original. No longitudinal studies of subjects' changes in performance or detailed reaction time analysis was done. Also, he did not measure reading times when subjects were given text, either for the entire text or for individual premises. This was probably partly due to technological limitations at the time of the experiment.

The type of model Johnson-Laird developed was determined by the combination of the constructivist theoretical position and the kind of data he collected. In the case of the spatial description domain, the model involved translating text into logical premises (at a local sentential level) and constructing an array representing the two dimensional layout of the objects in the description. The

statements are dealt with one by one, each being combined with the array constructed so far. If a statement cannot be incorporated into the array then it will have to be stored on one side, pending relevance. If this happens too much a memory overload will be created and the subject gives up the array and simply stores the set of premises. This is similar to the kind of thing that can happen with the 'working memory' of production system models.

The other example was syllogistic reasoning. How do subjects draw a 'spontaneous' conclusion from a pairs of premises such as the ones below?

Some of the artists are beekeepers.
All the beekeepers are chemists.

Conclude: Some of the artists are chemists.

This is example from the space of possible syllogisms formed by varying the quantifiers, which may be any of *all, some, some not* and *none.* Some combinations of premises are much more difficult to reason about than others (taking minutes, rather than seconds, for a naive subject), and are also prone to error. Part of the effect is syntactic, due to the order in which the premises appear, and Johnson-Laird calls this the 'figural effect'. However, some syllogisms are genuinely more difficult than others, even discounting this factor. Newell, amongst others, tried to model the reasoning processes involved but, as Johnson-Laird points out, he is unable to make specific predictions about errors. Although he used production systems for modelling, he was unable to use protocol analysis. Although subjects are able to report some aspects of their reasoning processes, such as the mental images they use, it seems that no one can give a step-by-step rationale.

Johnson-Laird's model involved subjects constructing and manipulating a kind of iconic 'tableau' as a result of reading the premises. The tableau would contain 'tokens' representing components of the problem, in this case, individual beekeepers or artists, or whatever. The idea is that, at any given stage, the tableau should be *representative* of the problem description, rather than *necessarily* following from it. A tentative conclusion is 'read off' the tableau and

then tested by manipulating the tableau, within the constraints specified by the premises, to see if it can be falsified. The tableaus constructed for some forms of syllogism tend to suggest erroneous conclusions which might not be spotted if the subject runs out of 'working memory space' whilst trying out alternative representations of the premises. This factor is used to account for subjects' errors.

For the spatial descriptions example, the spatial properties of the tableau (an array, in this case) are used to represent spatial relationships specified in a description. For example, the premise 'the spoon is to the left of the fork' would be provisionally represented by two tokens in adjacent cells in the array. Additional tokens are introduced into the array as further premises are encountered, with the array being rearranged, as necessary, to deal with inconsistencies. Questions about the description are answered by interrogating the array, to find the relative positions of the objects in question. If too many inconsistencies arise, however, (the description is ambiguous) subjects supposedly run out of working memory space and resort to remembering the descriptions 'verbatim'.

This account of understanding descriptions contains an implicit model of transitive inference. Formally, the reasoning involved in ordering objects in space is the same as involved in making the inference.

$$x > y \;\&\; y > z \rightarrow x > z$$

Yet Johnson-Laird is not concerned with the details of the inference mechanism. He does not specify how the tableau is 'read' and takes no account of the phenomena associated with transitive inference, discussed below. The same applies to Hagert (Hagert, 1984), who attempted to formalise Johnson-Laird's model. This shortcoming should not distract us from the value of Johnson-Laird's work. He is trying to set up a general framework, or metaphor, for understanding problem solving, not just a specific account of spatial or syllogistic reasoning. This, naturally, is a major undertaking, and consequently the mechanisms he proposes are bound to be under-specified with respect to the basic forms of inference.

2.3 Transitive Inference in Adults and in Children

The case was made in chapter 1 that transitive inference is a suitable domain of investigation. Transitive inference itself is an established sub-field of cognitive psychology and there are a number of robust phenomena associated with the field which are described below. However, the earlier studies do not provide us with a cohesive corpus of theory or data. This is because of (a) diversity in the motivation and theoretical approach of researchers and (b) diversity in methodology. For this reason, a historical perspective is adopted, initially.

2.3.1 The Three Term Series Problem

One of the difficulties is that there is a fundamental methodological problem in trying to assess whether subjects possess a conceptual understanding of transitivity. Piaget's initial work concentrated on obtaining verbal justifications for subjects' solutions to transitivity problems but, of course, verbal inability does not preclude a conceptual understanding or procedural type knowledge. All we have to go on is behaviour in tasks which appear to require the coordination of relational information.

The three term series problem, or 'linear syllogism' consists of two conjoined premises describing a ranking of three objects with respect to some scale. These are preceded or followed by a question about the ranking or a statement relating the end terms of the ranking to be verified or falsified. Early investigators of human performance on this kind of problem were Burt(1919), Piaget(1921, 1928), Hunter(1957), deSoto *et al*(1965), Huttenlocher(1968) and Clark(1969). Although the research interests behind the various research groups were very different, ranging from cognitive development to psycholinguistics, a major common finding emerged. This was that different wordings of the premises produced

radically different performances even though they appeared to contain, logically speaking, exactly the same information about a three term series. For example:

'If John is better than Pete and John is worse than Bill then who is best?'

Produces many more errors than

'If Bill is better than John and John is better than Pete then who is best?'

These examples are, subjectively, obviously different but there are also a range of intermediate cases which need explaining. The effect is entirely analogous to the 'figural effect' obtained with ordinary syllogisms, so it is referred to as that, hereafter.

Being syllogistic in form, these problems are immediately recognisable as examples of formal reasoning in the logical sense but are much easier (and faster), and so are suitable for younger subjects. The three-term series problem was later studied by Clark (Clark, 1969) who was a psycholinguist interested in how the meanings of the comparatives were represented and understood. As for the conventional syllogisms, the main form of data was the error rates for different forms of the problem. Clark wanted to know such things as whether 'bigger' is easier to understand than 'smaller'. Hunter (Hunter, 1957) on the other hand, was perhaps the first to worry about how the inference was 'done'. Although using similar data, he came up with the first theory that was applicable to series of more than three terms, and came closest to postulating some kind of internal representation of the *series* as opposed to the linguistic terms. Huttenlocher (Huttenlocher, 1968), Foos (Foos, 1980) etc extended the idea of a representation of a series further, postulating that the inference was mediated by a *mental line* or *image* or *map*.

2.3.2 The N-term Series Problem

The three term series paradigm is problematic in that although it is a rich source of phenomena, the various effects are easily confounded due to the limited nature of the task. The phenomena began to be factored out, and new ones discovered, when the paradigm was extended to 'N' term series, usually four, five or six terms long. The procedure is essentially the same as before: the subject is presented with a set of premises relating items adjacent in the series and is subsequently asked questions about remote pairs. This has a number of advantages over the three term series. Given the series $A > B > C > D > E$, there are not just one, but a whole set of possible inferences relating non-adjacent pairs, thus generating a much richer data set. Furthermore, in series longer than four items, it is possible to test subjects on pairs such as 'B-D' where neither is an end item of the series. This turns out to be important for eliminating possible 'non-logical' reasoning strategies involving labeling the end terms.

There are a number of ways the task can be presented. Usually the premises are presented individually, in randomised or particular orders, followed by a testing phase. Alternatively, the training phase can be like a task itself with the subject given forced choices between a pair of stimuli, immediately followed by feedback as to whether the choice is correct or incorrect. Testing on novel pairs starts once a criterion of so many 'correct' responses has been attained. This latter procedure has the advantage that it can be used to give a non-linguistic presentation suitable for experiments with young children or animals, as described in the following chapter.

2.3.3 Phenomena Associated with N-term Series Problem

This section describes some of the varied phenomena which have been associated with various forms of transitivity task. This summary collapses many versions of the task, linguistic and non-linguistic, child and adult, and covering widely differing experimental procedures. It is therefore unlikely that a single mechanism

is responsible for all the phenomena. However, there is a more uniform corpus of experiments employing more or less the same procedure, and these experiments are reviewed in the next chapter, under the heading of the N-term series *task*.

Notation

Before introducing the phenomena, it will be helpful to introduce some terminology. Given a series of symbols, for example, $ABCDE$, the items at the end of the series are sometimes referred to as *end-anchors*. The 'training pairs' are adjacent items in the series (*eg* BC) and 'remote pairs' consist of non-adjacent items (*eg* BE). The 'ordinal separation' of a pair is the 'distance' between two items with respect to an imaginary linear ordering, where the separation of adjacent pairs is 1. For example, the ordinal separation of BE is 3. The *ordinal position* of an item or a pair is its overall position with respect to the linear order (and a preferred or arbitrary direction). For example, the adjacent pairs can be ordered AB, BC, CD, DE.

Measures

The staple measures of the experimental psychologist are 'error rates' and 'reaction times' (RTs), although these are by no means the only possible measures. Studies in cognitive science usually make the assumption that differences in the reaction time involved in making a response reflect differences in amount of processing and hence the 'psychological complexity' of the problem. This is only true if the processing is serial (as with conventional digital computers) and if there is no 'speed-accuracy trade-off' affecting the reaction times for difficult problems. Nevertheless, the assumption is a useful one to make, especially if it is backed up by error rates data showing the same kind of pattern across problems.

Linear Spatial Images — evidence from protocols

When subjects are asked how they perform N-term series problems, they typically report forming some kind of image of the linear order with a spatial ex-

tent (deSoto *et al*, 1965; Huttenlocher, 1968; Foos, 1980). These images, mental lines, integrated representations, analogue devices or mental models, as different theorists have called them, have a number of properties, or constraints, not all of which are agreed upon. However, it appears that they are usually constructed either from left to right[1] or from top to bottom and that only one vector is used. In addition to questioning subjects, Huttenlocher gave subjects pen and paper for the duration of the task, finding additional evidence that subjects were compiling a linear order with a preferred direction of working. It has not been established whether or not such images perform a mediating role, as Kosslyn and others suggest, or whether they are an 'attendant' phenomenon without a functional role.

Ordinal Distance Effect

The most robust finding is that the reaction time to make a particular comparison[2] is linearly related to the ordinal separation of the two items being compared. The further apart the items are in the overall ranking, the shorter the reaction time is (Potts, 1972; Potts, 1974; Trabasso & Riley, 1975). Superimposed on this effect, is a time advantage to comparisons involving an end item (end anchor) but, significantly, there is a distance effect with the 'internal' items. This shows that the distance effect cannot be attributed to a statistical effect whereby remote pairs contain more end anchors (Sholz & Potts, 1974).

Trabasso and his colleagues attributed this distance effect to the properties of the 'mental line' hypothesised by the 'image' theorists. The main idea is that points on the line must first be located, and then 'discriminated'. The closer two points are the more 'confuseable' they are, and hence the longer it takes to decide which comes first or second. However, it does seem to be agreed that ordinal information is integrated by the subject and not stored as adjacent pairs

---

[1]For Western subjects, presumably.

[2]averaged across subjects and question types

(which would mean that a transitive inference rule would have to be used for queries about remote pairs).

The Acquisition Curve

During the training phase of the N-term series task (when subjects are responding to the adjacent pairs) the error rates and reaction times for each pair show an 'end anchor' effect. That is, to start with, performance is better on pairs which have been recently presented (McGonigle & Chalmers, 1984a), but as performance improves with further training, the profile changes such that pairs towards either end of the series have lower error rates and reaction times. Trabasso describes this as the series being learned 'end-inwards'. A graph of error rate against ordinal position thus typically shows an inverted 'U' shaped curve (Trabasso & Riley, 1975). Strong serial position effects, such as the acquisition curve, are symptomatic of the typical transitive inference task and it is sometimes difficult to disambiguate these and distance effects. Woocher (Woocher *et al*, 1978) has argued that nearly all of the variance in his data could be explained by serial position effects alone. However, the general consensus is that both types of effect exist.

The Bypass Effect

Polich and Potts (Polich & Potts, 1977) found another end-anchor effect whereby the distance effect appeared to be short-circuited when end anchors were involved. For example, given the sequence A to E, any comparison involving A would be equally fast, regardless of how widely separated the pair is. This could be explained if the subject aborted any further manipulation of his/her mental representation of the series once it was discovered that finding A effectively defines the choice. Polich and Potts noted that some of Trabasso's data showed the bypass effect (Polich & Potts, 1977).

Congruity Effects

Trabasso and his colleagues, working with children comparing coloured sticks of different lengths, found a kind of congruity effect (Trabasso & Riley, 1975) which may be related to the congruity effect found with memorial comparisons (section 2.3.5).

For comparisons including an end-anchor, there is an interaction between the comparative employed in the question (ie: 'choose longer' or 'choose shorter') and the end-anchor in the pair (the longest or the shortest item). In terms of the comparatives employed, the instruction 'choose longer' is 'congruent' with the 'long' end-anchor being involved, and the resulting reaction time is faster than if the anchor involved had been 'short' and the instruction therefore incongruent. The same principle predicts the interaction of the 'choose shorter' instruction with the presence or absence of the 'short' end anchor. From later work, it seems that congruity effects may not be limited to pairs involving end anchors (Woocher *et al*, 1978).

Marking

The marking effect has been given relatively little attention in the context of the N-term series task. It is an overall asymmetry in the speed with which questions involving one comparative are answered as opposed to its converse. For example, Trabasso found that, overall, 'choose longer' comparisons were faster than 'choose shorter'[3]. Marking is often associated with *lexical marking* (described below) which is a linguistic phenomena. However, marking effects can obtained even with non-linguistic mode of presentation of materials and questioning, which raises the possibility that marking is due to an attribute of a subject's representation itself and is not due to differences in processing time of the instructions *per se* (McGonigle & Chalmers, 1984b).

---

[3]As with the congruity effect, Trabasso unfortunately limits his description of the effect to pairs including end-anchors

If marking is due to a property of representation then it may be connected with an asymmetry noted in the protocols and in the experiments on imagery (Huttenlocher, 1968; Foos, 1980), which is in the *direction of working* when subjects mentally construct a series. This is in contradistinction to Trabasso's claim that subjects construct series 'ends-inwards', which he supports with the serial position effect. If our 'mental lines' are directional in some sense then the marking effect could be explained in a similar way to the congruity effect: instructions which match the overall direction of the series are processed more quickly. However, this apparently leaves us at a loss as to how to explain the cross-over effect.

Further investigation of the causes of marking are hampered in the transitive inference paradigm due to the difficulty of dissociating possible asymmetries in subjects' internal representations due to training-specific effects, more general asymmetries, and effects due to processing of comparatives. For example, task specific effects could occur in experiments where the premises are presented sequentially in an order reflecting the series being represented:

A is bigger than B, B is bigger than C, C is bigger than D etc.

Additionally, an asymmetry could arise if only one comparative is used in the training phase (McGonigle & Chalmers, 1984a). The assembly process is assumed to be directional either due to properties of the image itself (deSoto *et al*, 1965) or to properties of the comparative terms, for example, the subject-object relationship (Huttenlocher, 1968) or due to limitations of the assembly process (Hunter, 1957; Foos, 1980). Another possibility is that, even if both comparatives are used, subjects may tend to use a common direction of working due to linguistic or pre-linguistic conventions such as working from biggest to smallest, tallest to shortest, fastest to slowest etc. It is difficult to disambiguate these various possibilities from the transitive inference literature alone as there is a lack of reportage on the effect and a lack of concordance of methodology.

2.3.4 SDE Phenomena

Although they may be psychologically distinct, SDE phenomena have (historically) been closely linked to those of the N-term series, and this is the main reason for their inclusion.

Analogue and Propositional

Naturally, workers in the N-term series paradigm have tended to look to related fields in psychology for sources of theoretical insight. Most of the terminology arises from a mixture of related areas, and various researchers have turned to different areas for sources of analogy and explanation, depending on their theoretical predispositions.

Broadly speaking, theorists split into two camps, 'analogue' (imagery) theorists and psycholinguists or propositionalists. These camps approximately correspond to two sides of a long standing (sterile) debate in cognitive psychology over the role of 'analogue' devices or 'images' in mental representation. For a review of relevant literature see (Kosslyn, 1981) as example of an analogue theorist and (Pylyshyn, 1981) as an example of a 'propositionalist' who argues that all image-like phenomena can be explained by propositional (language-like or sub-linguistic) encoding that forms the basis of both verbal and non-verbal semantics.

The N-term series researchers have tended to be influenced by the 'imagist' way of thinking, particularly in their explanations of the distance effect which was the phenomenon which dominated early research. Interestingly, however, Trabasso cites Clark, a psycholinguist who did early work on the three term series, when referring to the congruity effect. Also the term 'marking' derives from 'lexical marking', another psycholinguistic phenomena. These terms are explained below.

The SDE paradigm

Symbolic comparison tasks involve use of 'transitive' comparatives in a different paradigm from the N-term series task. Here, understanding a comparative involves relating it to previously encoded information in long-term memory. Such tasks therefore, may not initially seem relevant to researchers interested in how transitive inferences are made or how a set of comparatives are integrated. This class of experiments must be considered however, as it potentially enables us to dissociate retrieval from encoding phenomena (in the sense of task-specific encoding). An example of such a retrieval operation is answering the question 'Which is bigger, a cow or a dog?'. 'Cow' and 'dog' are thus symbols, presuming that you know what they are, referring to an internal representation of the properties of cows and dogs. The comparative 'bigger' must somehow key into a particular aspect of the representation. This is distinct from *perceptual* comparison where attributes like size or speed can, loosely speaking, be observed and compared directly without recourse to long term memory.

Perceptual Distance Effect

This is a well established (*eg* Cattel, 1902) psychophysical phenomenon whereby the time taken to discriminate two objects with respect to some perceivable dimension is inversely proportional to the perceived difference between them (a logarithmic function of the physical difference). For example, the time taken to judge which of two objects is heavier is longer the more similar they are in weight.

The *S*ymbolic *D*istance *E*ffect

Moyer (Moyer, 1973) wrote a paper entitled 'Comparing objects in memory: Evidence suggesting an internal psychophysics'. The 'comparisons' involved were size comparisons of objects referred to only by name and the psychophysical effect obtained was that the discriminations were faster the bigger the difference in size

between the (imaginary) objects. This referred to as the *symbolic distance effect* or SDE.

The SDE has been found to be ubiquitous in memorial comparisons along virtually any kind of perceivable or imaginable dimension. Where the dimension is metrical (as opposed to ordinal) such as size, it seems that decision times are inversely proportional to the difference in subjects' subjective impression of size (or whatever) of the objects to be compared (Moyer, 1973; Banks *et al*, 1983). In other words, the bigger the difference in a subject's impressions of the sizes of two objects, the faster their decision as to which is bigger. Also, this metrical difference is a better predictor of reaction time than the ordinal separation (the difference in ranking of the objects with respect to the ranking of all the objects used in the experiment). For example, deciding which is bigger, a refrigerator or a walnut would be faster (on average) than comparing (say) a banana with a football. SDEs have been found with abstract scales such as 'pleasantness' and 'value' (money) (Pavio, 1975) and also with 'semantic' orderings such as the ranking of the words 'second', 'minute', 'hour'... 'century' (Holyoak & Walker, 1976).

2.3.5 Semantic Codes and the Linguistic Account of the Phenomena

Linguists became drawn into the transitivity task paradigm by way of the early three term series experiments alluded to at the beginning of this section. We now return to this, as an introduction to the linguistic viewpoint, which asks the question, 'How are comparatives understood?'.

As Johnson-Laird points out (Johnson-Laird, 1972), in a review of work on the three term series problem, these are not logically valid deductions without additional assumptions, or knowledge, about the properties of transitive relations. This knowledge can either be regarded as entirely part of our knowledge of language and the process of comprehension, as with Clark (Clark, 1969) and Johnson-Laird (Johnson-Laird, 1972), or as based on prelinguistic spatial knowledge (or other paralogical devices) onto which the terms

from our language map (deSoto *et al*, 1965; Huttenlocher, 1968; Foos, 1980; McGonigle & Chalmers, 1986). In either case, the linguistic component must include knowledge about which comparatives correspond to quantity differences and are therefore transitive, and what their converses are. As far as I know there has been no work on comparatives which are transitive and yet do not always map onto simple linear orders. Examples of these are 'inside of' and 'ancestor of', but further discussion of these is postponed until later.

Clark's Linguistic Account of the Figural Effects

With his linguistic account (Clark, 1969) argued that he could explain the figural effect with the aid of three psycholinguistic principles: the primacy of functional relations, lexical marking and congruence. These in turn rely implicitly on theories of decompositionality — the idea that complex terms can be broken down into linguistic primitives or 'base strings' (eg, Chomsky 1965). For example, *'John is worse than Pete'* breaks down into two strings (the functional relations) *'John is bad'* and *'Pete is bad'* conjoined in a comparative construction. The primacy of function relations principle states that these underlying base strings are more 'available' after comprehension than the comparative construction. In other words, it is easier to remember that John and Pete are bad than to recall that John is more bad than Pete.

The **Lexical Marking** principle deals with the asymmetry of opposites like *good* and *bad.* The idea is that one term is more complex and is in some sense a derivative of the other. The more complex term is considered 'marked' with respect to another if its meaning is represented in terms of the simpler 'unmarked' word. For example, *'birds'* might be encoded as the base string *'bird'* with a qualifying primitive indicating plurality. Comparatives involving unmarked base strings, hereafter known as unmarked comparatives, are supposed to be easier to store and retrieve than their marked counterparts.

It has been observed that most pairs of comparative and converse appear to be asymmetrical; the unmarked one tends to be neutral with respect to the

absolute properties of the objects with respect to the scale, whereas use of the marked comparatives tend to have the implicature that both of the items are ranked towards the marked end of the scale. The following real life example is from deSoto *et al*'s paper: A disappointed spectator at a baseball match says,

> 'I came to see which of you two guys is better- instead I'm seeing who is worse.'
> (deSoto *et al*, 1965)

Similarly, 'taller than' is neutral but 'Henry is shorter than Simon' suggests that both parties are on the short side An obvious criticism of this is that the chosen comparative also depends, pragmatically, on who or what the sentence is *about*. For example, if Henry and Simon are both tall but Henry is the topic of conversation then the statement 'Henry is shorter than Simon' is not at all unnatural. In this instance Henry is the topic and Simon is the referent for the comparison. However, this complication does not detract from the finding that, in general, unmarked comparatives such as bigger, better and faster are responded to more quickly than their marked counterparts and that it seems that, in natural usage, it is often the case that comparatives tend to be appropriate to the end of the scale that the compared items are located in.

Further evidence for lexical marking comes from developmental studies (Donaldson & Wales, 1970) which show that children learn to use unmarked terms before marked ones. For example, 'big' and 'not big' are acquired before 'small'. Also, the name for a scale is usually derived from the unmarked comparative, for example, length, width, depth, height etc (Johnson-Laird, 1972).

Congruency, in Clark's sense, is a principle governing the retrieval of information from memory. A piece of information can only be retrieved if it is congruent with the unknown being sought in terms of the underlying functional relations. For example, the question 'Who is best?' requires an X such that X is good. The question would thus not be congruent with a memorisation of the statement 'John is worse than Pete'. To answer correctly, in this case, requires an implicit reformulation of the question to 'Who is least bad' which matches the base strings of the stored information. This notion of congruence

is used to explain why questions which use a comparative opposite to the ones in the premises take longer and are more prone to produce errors. The analogous explanation for the 'image' theorists is that such questions have a direction opposite to that of the encoded representation so that, for example, if the 'direction' of the assembled series is best to worst (X is worse than Y and Y is worse than Z) then this not congruent with questions of the form 'Which is better?' (McGonigle & Chalmers, 1984a).

The relative difficulties of different wordings of the three term series problems can then be accounted for by mismatches between the directions of comparatives in the premises and the direction of working in assembling the series.

A Comparison of Imagistic and Linguistic Approaches

The advantage that the imagists have over simple linguistic coding models is that the latter does not even begin to explain how the inferences are made. As soon as the limited domain of the three term series is expanded to include more terms, it becomes particularly unclear how comparisons between items in the middle of the sequence (other than end items) can be made. For this reason, *BD* comparisons in the five-term series problem are regarded as crucial as they cannot be mediated by simple linguistic codes. The imagists, on the other hand were already treating the three term series as a special case of a more general type of problem. However, Clark's theory was able to correctly predict the results from negative forms of the premises such as 'John isn't as good as Pete', as well as having independent support for the lexical marking principle. The two approaches can be reconciled to a certain extent by taking a procedural or 'information processing' perspective and also taking into account differences in the tasks used by different researchers, such as whether the question is given before or after the premises are presented. For a more detailed treatment, the reader is referred to Johnson-Laird's review (Johnson-Laird, 1972). The three term series problem is not reviewed in more detail here as it is now considered too limited and prone to task specific routes to expertise, such as scanning the premises syntactically to find an end item matching the required answer.

Trabasso supposed that the distance effect he found was related to the symbolic distance effect, but this requires the assumption that ordinal comparisons are also mediated by a mental line. This has the attraction of bringing together perceptual, symbolic and ordinal comparisons into one framework with the 'mental line' as the unifying feature. However, this does not take into account the directional effects of **marking** or **congruity** mentioned in association with the three term series work and, of course, it does not address the problems of how or why a series is assembled in the first instance. Furthermore, it appears that there are important differences between perceptual and symbolic comparisons which strongly suggest that there is more to symbolic comparisons than an 'internal psychophysics' (Moyer, 1973). These findings are described below.

The Semantic Congruity Effect

This is an extension of Clark's concept of congruity described above. The phenomenon can be summarised as being an interaction between the overall (mean) position of a pair of items on a scale and the form of comparative used in a symbolic judgement. For example, (Shipley *et al*, 1945) found that subjects were faster at choosing the 'more preferred' of two colours when the colours had relatively highly rated preferences and were faster at choosing the less preferred when the colours had a low preference. The effect appears graphically as a cross-over if the reaction times for adjacent pairs are plotted against their positions in the scale. The 'congruity' then, is between the comparative and the location of the referrents on the continuum. Figure 2–1 shows a typical cross-over due to the congruity effect, such as obtained by (Banks & Flora, 1977).

The congruity phenomenon has been replicated by a number of experimenters with different types of comparative, including age, size, etc (see (Marschark & Paivio, 1981; Banks *et al*, 1983; McGonigle & Chalmers, 1984b) for reviews and recent experiments). There now appears to be a consensus among these researchers that the effect only occurs with 'symbolic' comparisons and not with perceptual ones (Marschark & Paivio, 1981). It seems that the stimuli must have a *semantic* significance for the subject *relevant to the task*,

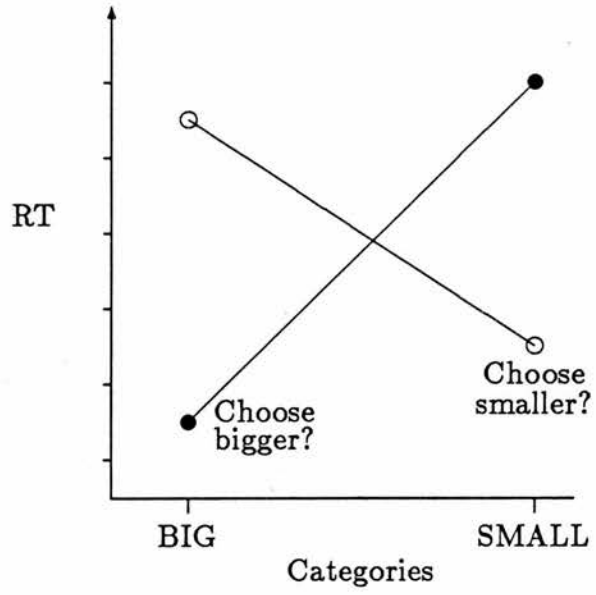


Figure 2–1: Stereotypical cross-over effect.

either by virtue of pre-existing knowledge of the stimuli involved or by knowledge gained through the course of the task — for example if only a small set of stimuli are involved and they are individually recognised.

As with the work on the N-term series, described in section 2.3.3, the kind of data and model generated in this paradigm has depended on whether a researcher been thinking in terms of (linguistic) comprehension of comparatives or in terms of analogical representations such as the mental line. Broadly speaking, each camp is able to explain different aspects of the phenomena but neither can give a full account inclusive of the other's paradigm. This situation is epitomised in the debate between Banks *et al* on the linguistic side, and Marschark and Pavio (Marschark & Paivio, 1981) on the analogical side. The main area of disagreement is over a phenomenon of *asymmetry* in the processing of comparatives: comparison is faster when the 'direction' of the comparative in the question (*eg* 'which is bigger?' *vs* 'which is smaller?') is in tune with the overall position of the items along the continuum (*eg* their overall size). This is called the *semantic congruity effect.*

Banks *et al* have typically done experiments involving their subjects getting

small numbers of items from a continuum (analogous to having a short series). On the basis of these they have argued that continua are effectively divided up so that sizes (or whatever) have a category coding, and it is this that is accessed (at least initially) with linguistic or abstract input. Thus there can be a congruency, in the sense of Clark, between the form of the comparative and the form in which the data is stored. Marschark and Pavio, on the other hand, have typically used a bigger set of sizes for each subject. Their account of semantic congruity is based on *expectancy* or *priming* in which the form of a comparative gives rise to a facilitation of dealing with a particular portion of the range (or mental line). Thus the question 'Who is better player, Fred or Jim?' would lead to an expectancy that both players will be at the good end of the range of players, and response will be slower if this is not the case.

A way of generalising the emergence of the congruity effect is that it increases with reduced response uncertainty due to knowledge about how the stimulus variation is constrained. For example, Marschark and Pavio suggested that Audly and Wallice obtained a congruity effect with brightness discriminations (which would normally count as perceptual discrimination) due to the fact that they used only two different stimulus pairs with a highly repetitious task. In this instance, the emergence of a congruity effect could be taken as indicative of *expertise* in the task. A different type of knowledge was important in a task originally invented by Banks *et al* but replicated by Marschark and Pavio with an interesting variation. The task was to choose the higher or lower of two circles above a line. However, some subjects were additionally told that these represented helium filled balloons floating on strings (tied to the line). Not only was this group relatively faster, but it showed a significant cross-over (congruity) effect which the other 'perceptual' group did not. Presumably, the effect of the world knowledge brought to bear by these subjects was to highly constrain the expected stimulus variation, thus making the task easier.

Banks *et al*'s interpretation of semantic congruity is that the internal representation of stimuli is not merely analogue but contains a categorical element in the form of semantic codes (as with Clark's model). According to Banks posi-

tion, there is no other way to explain the cross-over effect; and lexical marking and semantic congruity go hand in hand — congruity due to matching (response stage), distance due to discrimination stage.

Marschark and Pavio, on the other hand, argued that marking and congruity effects are mutually exclusive, congruity effects only being obtained when comparatives precede stimuli in the subjects' task. Their model of congruity is that of an *expectancy* being set up by the comparative (eg. choose bigger) for a particular part of the range of stimuli (eg. the larger objects in the set of stimuli).

There are a number of problems with this research which we will only touch on briefly here. Firstly, neither side deals effectively with the crucial issue of how a range of values is established by the subject as being relevant to the task. Presumably, without a range being established then the semantic congruity effect could not occur. Secondly, neither side has really gone into the information processing aspects of information retrieval and the comparison process, their models being essentially psychological (descriptive) in character. Finally, there is some difficulty relating the SDE findings with the N-term series task because of the extra layers of complexity introduced by the fact that information must be retrieved from long-term memory and because of the variety of testing procedures used. In many ways the N-term series task is a more suitable for modelling purposes, but providing a comprehensive account of various forms of distance effect is an important topic for future research.

2.3.6 Symbolic Comparisons: Conclusions

In both the N-term series and the SDE paradigm, the kind of data and model generated has tended to depend on whether the researcher has been thinking in terms of (linguistic) comprehension of comparatives or in terms of analogical representations such as the mental line. Because of these differing motives, the design of experiments has also varied, to such an extent that it is difficult to sort out which phenomena are associated with the *representation* of order, which are associated with linguistic access or encoding, and which are due to the effects

of contextual knowledge. It would seem that the SDE and related phenomena are not likely to succumb to a simple kind of model, especially as the literature seems so confused. First, the modeller would have to decide whether to tackle the representations of magnitudes in long-term memory or just a 'working memory' representation sufficient for the task demands. Many of the researchers, appear to assume that their tasks are performed by accessing long-term memory directly, but this seems unlikely where many comparisons are made involving the same set of stimuli. The processes which it might be necessary to consider are therefore:

1. The representation of sizes (or whatever) in long-term memory.

2. The affects of query form on retrieval from long-term memory.

3. The effect of the task context on the transfer of information from long-term memory to working memory.

4. The form of representation of sizes in working memory.

5. The affects of query form on retrieval from working memory.

Given the large number of factors at work, the lack of consensus in the SDE field is not too surprising. However, the N-term series problem seems more tractable. The N-term series problem also appears to lie at an interesting intersection between two disciplines. On one hand there is a set of theories and experimental data which conceptualises the act of making comparisons as a kind of symbolic extension to perceptual comparisons. On the other hand we have a body of literature treating comparatives as composite linguistic objects which are matched against similar objects stored in semantic memory. Neither approach gives a satisfactory account of all the phenomena. Further lacking, is an explanation of *how* comparatives are dealt with in terms of *process*. This paradigm seems ripe for the application of AI techniques, both for analysis and as a source of new metaphors, but what methodology should be employed? Johnson-Laird's work, described previously, employs some computational metaphors but his approach

is essentially psychological in character. Protocol analysis, described below, integrates computational and psychological methodology in a more intimate way.

2.4 Protocols and State Space Search Models

This approach stems from the work of Newell and Simon (Newell, 1977) who attempted to provide a general computational account of human problem solving using the concepts of state space search and means-end analysis. They started with the domain of 'crypt-arithmetic' in which adult subjects were given complex puzzles where they had to crack a cipher in which letters of the alphabet stood for digits in an arithmetic sum. They analysed the task in terms of sub-problems in association with getting the subjects to give detailed verbal accounts of their steps towards a solution and recording all the partial solution states. Armed with the concept of 'state space search', from computer science, they were able to think of subjects' strategies in terms of paths through a space of possible partial solution states. The states were linked by 'operators' such as processing a particular column or hypothesising a particular number for a letter. It is not clear how Newell and Simon arrived at the set of operators that they used to represent subjects problem spaces. Presumably it was a mixture of a programmer's intuition about how to solve the problem and examination of the verbal protocols. Newell simply states: 'The problem space is a hypothesis about the subject's behaviour'.

Following this stage, called a *task analysis*, an individual subject's path through the problem space would be mapped. The next step is to try and capture any regularities in these graphs in a *production system*, a high level computer language (explained below).

A major problem with this and subsequent work in this vein (*eg* the work on the 'Towers of Hanoi' problem) is that there is a lack of understanding as to how subjects actually create the search space for themselves. Although the experimenter can think of the subject as exploring a pre-existing space of pos-

sibilities, in reality the space must be discovered dynamically. A more serious limitation for our purposes, as already indicated, is that the technique is not readily applicable to fundamental forms of inference.

Production Systems

A 'production', or 'production rule' (Newell, 1973) is simply a condition-action pair which can be read as an *'if... then... else...'* statement:

If (stimulus) *then do* (action) *else* (return control).

Which can be denoted by the shorthand: 'Stimulus ⇒ Action'. Therefore, if procedures are seen as consisting of sequences of *'if... then... else...'* decisions, then production rules are a convenient notation. Notice that the *else* part of the statement contains the instruction to 'return control'. What happens in working production systems is that at this point control is returned to the system, which needs (itself) to have a decision procedure for selecting the next rule to try. For example, the simplest type of control is to order the set of rules in a stack and, starting at the top, test the condition of each rule in turn until one matches the stimuli. When this happens we say that the rule 'fires'. The other aspect of control is what happens *after* an action is carried out, for example, whether the control process stops, continues where it left off or starts again. This aspect turns out to be important and is returned to again in chapter 4.

In more complex production systems the conditions and actions of production systems can refer to and manipulate a global database. Such systems have full Turing power (they can mimic any process that a serial computer can carry out) and thus, in themselves, constitute no more than a notational system, in the same way that any programming language is a notation for procedures. They become psychologically interesting only when sufficiently constrained to make clear predictions. For example, the nature of the stimulus and/or the action can be restricted.

It should also be noted that there exists another aspect of production systems work which is more to do with modelling human memory than with problem solving behaviour. An analogy is often drawn between human *short-term memory* (as discussed by psychologists) and *working memory* in production systems. *Long-term memory* is a more permanent record of world knowledge and this is a whole topic of study in itself. This work is not relevant to this thesis, although ideas about working memory are important for some models of relatively complex forms of problem solving, such as described by (Johnson-Laird, 1983).

2.4.1 Seriation

Young's work on seriation (Young, 1976) is an example of a production systems approach to a simple Piagetian task. It is explored in some detail because of its relevance to transitivity and because it typifies the methodology. In a typical version of the task a child must form a line of wooden blocks of descending size by repositioning them from a scrambled 'pool' at a separate location. The overall search space for this task is surprisingly large and messy when one considers each operation at a fine grain of analysis and the redundancy in the possible operators which may be applied at any stage (think of the number of sorting algorithms in common use in programming).

Previous analysis of the task had been according to Piagetian *stages* on the way to fully *operational* problem solving. Piaget's observations (Piaget & Inhelder, 1969) were motivated by a coarsely *descriptive* theory so he necessarily overlooked many of the subtler (though important) variations in behaviour. In contrast, Young, with his finer grained analysis of the *processes* involved in seriation, was able to see that Piaget's *stages* are inadequate as a means of chunking the behaviour.

Piaget's Phases

Piaget's 'genetic epistemology' categorises children's intellectual development into discrete *phases.* Performance on tasks involving 'concrete-operational' skills

(direct operations on the world as opposed to entirely abstract or 'internalised' reasoning) can be further subdivided into three *stages* of competence. A child at *stage I* can perform some of the required individual operations (having the required physical skills) but not the full task. At *stage II* problem solving behaviour is *empirical.* This essentially means that the solution of a task is recognised but the operations needed to bring about the required state must be arrived at by trial and error. At *stage III* performance becomes *operational*; the child is capable of constructing an algorithm for doing the task.

The seriation task was one of those used by Piaget to help define and test his categorisation described above. The subject must physically sort a jumbled set of blocks of different sizes. The child is instructed to take blocks from the *pool* and assemble them in a *line* of decreasing size at a nearby, but distinct location on the table. Young's detailed analysis of a number of children's performances on this task showed that Piaget's crude characterisation of stages overlooks a rich variety of problem solving behaviour.

Methodology

The components of Young's analysis were as follows:

1. A task analysis (finding what processes will lead to a solution).

2. An analysis of subjects' behaviour (finding how a subject's strategy decomposes). This may be tested by experimentation.

These analyses are not independent, of course. The description of subjects' behaviour is in terms of the task analysis, and the task analysis draws on the description of subjects behaviour. This process results in a procedural model of a subject's problem solving strategy. Young refers to the interactive process as *adaptive experimentation.* and regarded production systems as '... an essential complement to the technique of adaptive experimentation...'. There is no doubt that production systems provide a useful notational medium for this kind of

model development, but the use of alternative procedural representations should not be precluded, as is discussed in a later section.

There is a danger that the power of 'adaptive experimentation' in developing procedural models may be abused. The technique, in itself, does not enforce the distinction between the *evaluation* of a model and its *development.* That is, it may become unclear where the model is being informed by observations as opposed to being tested, in which case the model might not be falsifiable. Young counters this problem by working out predictions of the models in advance of collecting new data.

Preliminary Task Analysis

The iterative process of developing a model starts with a 'preliminary task analysis' which decomposes the task according to a combination of intuition, preliminary observation and logical analysis of the steps necessary to allow successful performance. At this stage, Young hypothesises that the locations of blocks can be chunked into the *pool* and the *line* and that the seriation process can be meaningfully chunked into *episodes* consisting of the transfer of a block from one location to the other. Now the behaviour can begin to be described. An episode (usually) involves scanning the pool, reaching towards a block, grasping it, placing it on the right of the growing line, examining the line, possibly switching blocks round, etc..

It may seem that the basic categories (italicised above) are obvious and that it is difficult to imagine any other chunking, but even this stage introduces assumptions into the theory. An example of this is that the *episodes* consist of the selection, evaluation and *final placement* of a block:

Selection: the choice of which block to work with next. The basis for this may be the convenience of its location in the pool or having a particular block 'in mind' because it has been dealt with recently.

Evaluation: the decision whether or not to accept a block as suitable for inclusion in the line (at its current stage). Possible evaluation criteria are:

- similarity in size to last block in line (with some threshold)
- overall bigness
- nil evaluation (accept any selected block)

An evaluated block may be referred to as *oversize* or *undersize.*

Placement: whereabouts in the line growing line a block should be put. The default placement is at the right of the line, but this may be corrected by *switching* a block with its left hand neighbour. *Insertion* is a special form of placement where a block is directly inserted in its appropriate position.

Evaluation and placement are themselves complex processes. Evaluation may involve rejection of a block (if it fails some criterion), and reselection. Placement involves assessment and possible repositioning. The process of chunking is informed by observation but is not straight-forward. The categorisations above nevertheless form viable *working assumptions.*

To illustrate the point that alternative chunkings are possible, an alternative *episode* would be one beginning with the selection of a block to work with, and ending with putting the block down in one of the possible locations. This allows for the possibility of an episode ending in rejection of the block and its return to the pool. It could also be argued that there is a case for identifying an extra location of the table in which rejected blocks are placed (as distinct from the pool). This would also have been consistent with the preliminary observations. However, minor problems with the preliminary analysis can be corrected during the subsequent process of 'adaptive experimentation'.

Protocols

Once a descriptive language has been developed, videotapes of subjects' performances can be transcribed onto paper in the form of a 'blow by blow account'.

Episode	Summary	Line
1. Add F	Scan Pool Reach towards E Get F, Put at left	 F
2. Add C	Get C, put next to F Examine	F C
3. Add E	Get E, Put next to C Examine Switch C, E Examine	F C E F E C
4. Add B	Get B, put next to C Examine	F E C B
	etc	

Table 2–1: Sample from a protocol of a child seriating (from Young).

Protocols are often described as the starting point of the modelling exercise, but the preceding observations should make it apparent that this is not strictly the case. Table 2–1 contains a sample from a protocol, showing the level of detail involved.

A production System for Seriating

Young claims to capture the essential features of this behaviour in a simple production system. This contains rules like the following:

Goal=Seriate ⇒ Set.goal[ADD.ONE]

Goal=ADD.ONE ⇒ Get.block[next nearest]

Goal=ADD.ONE and have.just [Get.block'd] ⇒ Change.goal.to[PLACE]

Goal=PLACE ⇒ Put.block.at[right]

Goal=PLACE and have.new.configuration ⇒ Examine

⋮

Unfortunately, Young does not give a full description of the action of the interpreter except to say that the conflict resolution rule is to pick the rule with the 'most restrictive' preconditions. This appears to mean that the rule with the most (true) preconditions is the one that fires. The notation for the rules is clearly fairly informal. For example, he does not explain how the preconditions "have.just" and "have.new.configuration" operate. Clearly, they refer to internal states and not to perceivable features of the task, but how and when is the appropriate information stored? Does the system have access to previous states arbitrarily (psychologically and computationally implausible) or is it the case that only operations or outcomes that are known to be needed in the future are stored? Also, the treatment of sub-goaling is confusing, as it is not clear what happens when goals fail.

Yet, informal as the procedural notation was, it clearly assisted Young in gaining insights into the strategies employed by his subjects. The modularity of the representation was, in part, successfully matched up to the modularity of the behaviour, enabling commonalities and differences between subjects of different ages to be described.

In summary, whereas Piaget was drawn to the repetitive or *algorithmic* aspects of able subjects' behaviour, Young noted the *flexibility* of subjects' responses in the way they adapted to local conditions during the problem solving process. He captured some of this flexibility by representing subjects strategies as production systems with some redundancy (overlap) in the conditions of different rules. The *group* problem space was represented by a 'kit', or superset, of production rules out of which subsets were selected to model individuals, including 'non-seriators'. He was thus able to create hypotheses about what procedures subjects held in common and to account for differences in style between subjects of the same ability. Whilst he was not able to provide an account of how subjects might come to acquire rules of the type in his model, he did show

how a progression through different stages of ability could be modelled by the addition of extra rules onto a core set.

2.4.2 Protocol Analysis: Conclusions

The previous examples serve to illustrate how a combination of task analysis in terms of a state space search, protocol taking and production systems can lead to a particular kind of model of problem solving which is, in many ways, more satisfactory than earlier descriptive psychological accounts. The main limitations of this methodology are as follows:

1. It may not be clear what the appropriate task analysis is.

2. There may be difficulties in matching the level of descriptions in protocols with the level of the task analysis.

3. There is no way of modelling dynamic changes in the search space due to the learning of new operators or alterations to existing operators.

However, there are a number of useful lessons to be learned:

1. The psychological evidence must be collected with microanalysis and procedural modelling in mind. For example, Young had to collect his own data based on in-depth 'longitudinal' studies on a few subjects. This was despite many previous experiments having been carried out on seriation. The Piagetian school have tended to use group data for analysis, for the purpose of descriptive classification of behaviour.

2. The assumptions behind the task analysis need to be made explicit, in case it has to be revised.

3. The procedural implications of a model need to be made as explicit as possible.

4. There is a need to separate out the iterative stages of model development and evaluation.

2.4.3 Selecting a representational form

If subjects' procedures are modelled by a rule based system of some kind, this amounts to assuming that it is possible to represent the underlying reasoning processes by the logical manipulation of discrete symbols. This assumption is rarely made explicit. A legitimate alternative approach would be to try and use a parallel-distributed processing (PDP) model. The advantage rule based systems have is that they are relatively well understood in AI (more research effort has been invested in them than PDP systems) and there are a host of techniques available for representing determinate, sequential symbolic processes. The cost of making this simplifying assumption is that whilst we may be able to reflect the logic of what is going on inside subjects' heads, the actual implementation of the decision processes may involve parallel distributed processes. If this is the case then predictions about the *process* (*eg* reaction time) are likely to have limited success, as complexity does not bear the same relationship to processing time in sequential and PDP systems.

The next step is to select a suitable formalism for symbolic manipulation. Mathematical logic (the theorem proving metaphor) would appear to be the obvious choice as it is the only currently available 'common currency' for comparing different representational forms and is completely general. However, providing a procedural interpretation for a set of logical formulae is not always straightforward. Moreover, logical notation is not yet generally adopted by the psychological and cognitive modelling communities so there is an advantage in using a simpler (and less general) notation if possible. Production systems, in their basic form, are such a notation and have the advantage of being fairly intuitive, especially to anyone with a knowledge of conventional programming languages. However, it is sometimes useful to represent the final model in alternative notations, such as mathematical logic to enable comparison with other models. This approach is adopted in chapter 6 for comparing models of transitive inference.

2.5 Overall Conclusions

A number of attempts to understand or model problems solving and inference have been reviewed, with the emphasis on studies related in some way to transitive inference. Existing models have tended to fall into the following categories.

1. Those that have embraced the computational metaphor, and tried to give *procedural* accounts of reasoning, have concentrated on relatively high level reasoning where the intermediate states of problem solving can be intuitively and/or empirically identified.

2. Researchers coming form a background in psycholinguistics have been concerned with the representation of the meaning of individual sentences involving comparatives rather than the processes of inference.

3. Some models rely heavily on an intuitive notion of 'mental image' without specifying the computational processes that this implies.

Johnson-Laird's work appears to resist such a straightforward classification, but on closer examination, it can be seen that his models are composites, rather than truly crossing boundaries. For example, his model of the comprehension of spatial descriptions uses both an 'image' and 'superficial linguistic representations' to explain subjects behaviour. Although the existing methodologies fall short of providing adequate explanations of fundamental forms of inference, there is a lot to be learned from previous modelling attempts, and existing methodology can be built upon.

In particular, previous psychological studies of transitive inference have failed to give satisfactory accounts, and this may be partly due to the underestimation of the complexity of the problems involved. Although transitive reasoning superficially appears to be a simple form of reasoning, it can appear in several guises and its operation can be difficult to disentangle from that of linguistic and memorial processes. Just as data needs to be collected explicitly for the purposes

of protocol analysis, there is a need for in-depth, long term studies of transitive inference, if it is to be adequately modelled and not simply described. Studies of the N-term series task, described in the following chapter, come closest to satisfying these requirements, and so this is the best domain to focus on.

Chapter 3

Non-Verbal Transitive Inference

The main purpose of this chapter is to introduce the data base for the modelling work presented in subsequent chapters. Before this however, some studies of the N-term series task in children and birds are reviewed. Although the data from these studies does not meet our modelling requirements, the same basic experimental paradigm is employed, and the common findings demonstrate that the form of reasoning to be modelled is not species specific. Tables and figures in the monkey section are provided courtesy of McGonigle and Chalmers.

3.1 Children

This literature on transitive inference in children has been extensively reviewed elsewhere (Trabasso *et al*, 1975; Breslow, 1981; McGonigle & Chalmers, 1984a) & (McGonigle & Chalmers, 1986), so only an overview is given here. Bryant and Trabasso developed new versions of the transitive inference task for two main reasons. The first was that they challenged the accepted wisdom of the Piagetian school of developmental psychology that young children do not have the logical mechanism which is needed to co-ordinate separate items of information in an inference. They hypothesised that the apparent inability observed by others, *eg* (Piaget & Inhelder, 1969), was due to lack of retention or understanding of the original premises rather than inability to put them together. The second reason was their concern that where subjects (older children) did appear to show an ability to coordinate relations, this might be due to 'parroting' a verbal label picked up in the initial training. For example, the child could represent '*A*

is bigger than B' using categorical labels such as '$big(A)$' and '$small(B)$'. A second relation, 'B is bigger than C', could be represented similarly as '$big(B)$' and '$small(C)$'. In attempting to combine this information, the child is left with only two unambiguously labeled objects, A and C, with B being both 'big' and 'small'. In comparing A and C all the subject has to do is recover their respective categories.

Although both the above criticisms of the three-term series task had been raised before (Smedslund, 1966), they had not been dealt with within one experiment. Bryant and Trabasso's solution to the first problem was simply to ensure that the children had a lot of experience with the premises (the initial comparisons), and then to test them for recall of the premises at the time of testing on the inferential comparisons.

Their control for the possibility of 'parroting' was to extend the series to five terms. If the above strategy is applied to five terms (four pairs), then it only works for comparisons involving an unambiguously labeled item at one of the two ends of the series. Given the series $A > B > C > D > E$, the 'internal' items B, C and D, will end up being simultaneously labeled as 'big' and 'small'. The five-term series thus affords one critical comparison, 'B *vs* D', which is neither one of the original premises (the adjacent pairs), nor contains an item which could be unambiguously labeled. A six-term series would afford three such comparisons.

Another, related, argument for using such a test is that ability to make a 'B *vs* D' comparison would rule out a simple 'associationist' (stimulus-response) learning theory as being applicable to this task. In a five-term series task, the objects B, C and D are associated equally with selection or rejection and so there is no grounds for discriminating between them. A more sophisticated 'connectionist' type model would be needed whereby objects become associated with each other and there is a process of 'generalisation' of association (see Trabasso's model, chapter 8).

Bryant summed up the necessary experiment as follows:

> 'The correct way to test for inferences in young children, therefore, is to have four initial direct comparisons, to make sure that the child

> knows these fairly thoroughly, to test the child's memory for them at the same time as testing his ability to combine them inferentially, and to make the BD comparison the crucial test of the child's ability to make inferences.' (Bryant, 1974)

The first experiment along these lines (Bryant & Trabasso, 1971) involved five coloured sticks of different lengths as stimuli. The subjects were in three groups, with four, five and six year old children respectively. In the 'training phase' the sticks were presented in four pairs $A > B$, $B > C$, $C > D$ and $D > E$ where the ranking of the sticks from longest to shortest was $A > B > C > D > E$. The procedure was to present a pair of sticks sticking out of a block of wood with only an inch of each stick showing. The subject was then asked which was taller (or shorter) and then shown the whole lengths of the sticks afterwards, by way of feedback. The same pair would then be repeatedly presented until the subject had learned to identify the correct answer by the colour cues alone. The same procedure was then repeated for the other three pairs of sticks in turn. The second phase of training involved presenting a different pair on each trial, until the subjects reached a criterion of 90% correct on every pair. The testing phase then began.

The subjects were tested on all ten pairs derivable from the five-term series; the four original pairs, the critical BD pair and five pairs involving one or both of the end terms, A and E. No feedback was given on any of these test trials. The percentage of transitive responses on the BD pair were 78%, 88% and 92% for 4, 5 and 6 year old children respectively; results which are all significantly above chance levels. Responses on the other pairs were all similar or better. Interestingly, the proportions of correct responses on all pairs involving an end-term were generally higher than for the training pairs BC and CD. This suggests that labeling of end-terms could play a significant role and so the precaution of using a five-term series is a valid one.

Bryant and Trabasso concluded that children could indeed make deductive inferences and that Piaget and Smedslund were too 'pessimistic' about their abilities. They reasoned that retention of the original premises was the crucial

factor and that, furthermore, errors on the BD pair were probably due to lapses of memory for the BC and CD premises.

There remained the possibility, however, that subjects were performing the task (in whole or in part) by remembering the absolute lengths of individual sticks, which ranged from seven inches to three inches long in steps of one inch. A second experiment was conducted to eliminate this possibility in which visual feedback was not given in the training phase; subjects were simply told whether they were right or wrong. This made the training phase more difficult, but otherwise the results were similar to those obtained previously. See (McGonigle & Chalmers, 1984a) for discussion about the role of feedback during training.

3.1.1 Conclusions

Although Bryant and Trabasso developed a very promising paradigm for inference research, they did not do in depth longitudinal studies on individuals nor any kind of post-test which would give further insight into the mechanism employed.

3.2 Birds

Von Fersen, a student of Professor Delius[1] has recently carried out a five-term series experiment on pigeons. The research is still in progress but the preliminary findings are described here.

The basic training procedure was similar to that described for the monkey work below except that the stimuli were five blacked-in, irregular shapes with

[1]Dept. Psychologisches Institut, Ruhr Universitat, Bochum, W. Germany (personal communication)

rounded contours, instead of colours, and the subjects pecked keys instead of displacing tins. Pairs were presented in random order throughout the experiment except for a few session in which they were presented in blocks. On the crucial *BD* pair, subjects chose transitively on 87% of trials. This level of transitivity compares favourably with previous studies.

The only additional point of interest here is that the pigeons appeared to employ the same representational strategy even when they were offered a simpler alternative. The subjects were divided into two groups. The first were presented with a training set in which the stimuli shapes were of approximately equal area, but for the second group, the same five shapes were graded in size from one end of the series to the other. The ranking by size reflected the symbolic ordering. It was expected that the ordering of the stimuli according to a perceivable dimension would either facilitate the learning of the symbolic series or perhaps be used directly as the basis of discrimination. However, it turned out that the difference between the two groups was negligible over all but the first few trials. This would appear to suggest that subjects do not employ an analogue representation such as the 'mental line' in this task.

3.3 Monkeys

McGonigle and Chalmers adapted Bryant and Trabasso's task for use with monkeys and their results were published in an article in *Nature* (McGonigle & Chalmers, 1977). The shift to non-verbal subjects was interesting for a number of reasons.

1. It was not known whether non-human primates could do an abstract reasoning task such as this.

2. If they could perform the task it was of interest to study the process in non-linguistic subjects so that it could be established whether the processes

involved were pre-linguistic (or, at least, not dependent on a language faculty). One tenable hypothesis is that inference is actually founded on such pre-linguistic structures rather than deductive reasoning being based on linguistic abilities (McGonigle & Chalmers, 1986). This idea is also compatible with Bryant's position (Bryant, 1974) that deductive abilities are present in very young children and that they underpin perceptual learning.

3. Animal subjects could be tested more intensively and over a longer period, allowing high density data to be collected with the potential for follow-up studies on the same subjects (children, on the other hand, are more difficult to motivate in such a repetitious task). This is a very important point from the perspective of modelling the data.

3.3.1 First Study

The original procedure is described in (McGonigle & Chalmers, 1977). Eight adult squirrel monkeys were presented (in a Wisconsin General Testing Apparatus) with a series of choices, each between two tins differing in colour. Each subject was presented with four pairs of colours, first in sequence and, later, in random order. On each trial, the subject's task was to displace one of the tins. 'Correct' choices were rewarded by the discovery of a peanut hidden under the appropriate tin. Displacement of the wrong coloured tin produced no reward. Colours were drawn from the set yellow, blue, green red and white and different subjects were given different combinations of colour and reward to counterbalance for absolute colour preferences. Furthermore, the left-right location of reward was randomised. However, for individual subjects the pattern of reward and non-reward was consistent according to the following schema, with the letters A to E identifying five colours:

A	*B*	*B*	*C*	*C*	*D*	*D*	*E*	Pair
0	+	0	+	0	+	0	+	Reward
Light	*Heavy*	*Light*	*Heavy*	*Light*	*Heavy*	*Light*	*Heavy*	4 subjects
Heavy	*Light*	*Heavy*	*Light*	*Heavy*	*Light*	*Heavy*	*Light*	4 subjects

The first two rows show how the four pairs were arranged to be adjacent in a five term series. The second two rows refer to a second type of feedback that was given to subjects during training. One of the tins in each pair was weighted so that for half the subjects the 'correct' tin was lighter and for the other half the correct choice was lighter. The idea behind this was to facilitate the formation of an ordering on the items (from heavy to light or *vice versa*) by the subjects, even though only two weight values were used. This is analogous to the procedure used by Bryant and Trabasso in which coloured rods differed in length but the difference could not be perceived by the subject until after the choice had been made. However, in their second experiment (Bryant & Trabasso, 1971), subjects were given no feedback at all as to the actual length of the rods and there was no essential difference in the results apart from the training taking longer. There is no reason to suppose that the use of binary weight differences in the monkey experiment makes a qualitative difference to subjects' learning or performance over and above the effect of underlining the feedback given by presence or absence of reward.

When the subjects had reached a performance criterion of 90% correct responses they were transferred on to a random sequence of training pairs. After this they were tested on novel pairs, including the 'critical' BD pair. No more than two test trials were administered in a session of ten and a session was not begun until the subjects had attained a performance criterion of 22 correct responses in the course of 24 successive, randomly ordered trials. One subject failed to meet these conditions and was rejected. During testing sessions, all choices were rewarded. The arrangement of the testing sessions was thus designed to measure any spontaneous bias that the subjects might have on novel pairs.

The overall result (details are given further on) was that subjects responded with a transitive bias to all ten pairs under these test conditions. A second kind of test was then given in which the subjects were presented with triplets from the series instead of pairs. These test 'triads' were administered in an analogous way to the pairs, with subjects being rewarded whatever they chose. The motivation for these tests was to see whether or not presenting more items from the series would facilitate performance. For example, with the crucial comparison B *v* D, a 'coordination' type of model (Bryant & Trabasso, 1971) would predict that the middle item, C, would have to be retrieved as a referent in order to mediate the comparison. If this were the case, then presenting all three items explicitly in the triad BCD ought to facilitate the inference. Instead, performance on the triads was worse than on the pairs, with many choices going to the 'middle' item (C in the previous example).

The entire experiment was repeated (over six years later) in a second study with five of the original subjects being retrained on the same series as they had previously learned. In the Second study, however, reaction time (RT) measures were taken and along with video recordings. RT and video protocols are dealt with in a later section. The second study also featured an additional, more intensive phase of testing on the triads. In the second study, *no further training or selective feedback was given once testing had started.* So as not to confuse the different sets of data, the following naming conventions will be employed.

3.3.2 Summary of notation

The following notation has been adopted for convenience of reference in the modelling chapter. Where it differs from McGonigle and Chalmer's notation, the alternatives are shown in brackets.

1. The following data sets are referred to:

   (a) **Study 1** (1977, or 'original'). This includes binary and triadic choice profiles but no reaction times. The triadic tests are referred to as

early. The seven subjects' names were Bill, Bump, Brown, Blue, Roger, Green and White.

(b) **Study 2** (1983). Brown, Blue, Roger, Green and White were re-trained. Following this subjects were given, in chronological order:

i. **Binary RTs** — Reaction times taken during binary tests.

ii. **Middle** ('shallow' or 'early') — initial triadic tests.

iii. **Late** ('dense') — intensive triadic tests.

There are thus three phases of triadic tests, *original*, *middle* and *late*, with re-training coming before the latter two.

2. Subjects were trained on adjacent pairs in the series A, B, C, D, E, with items towards the E end being rewarded. Responses to untrained pairs or triads are deemed to be *transitive* (correct) if the item nearest the E end is selected.

3. The following convention has been adopted when referring to items in triads, for the purposes of clarity. The item nearest the E anchor is referred to as α[2] (the 'transitive' choice), the middle item as β and the 'worst' choice (the item nearest the A anchor) as γ.

	A	B	D
Eg:	γ	β	α
	(Non-transitive)	(Intransitive)	(Transitive)

[2]Note that α is chosen as a symbol because it stands for 'best' and should not be confused with A.

	B	C	D	E
A	98	100	100	100
B	–	93	90	76
C	–	–	89	87
D	–	–	–	97

Table 3–1: Percentage transitive choices on all pairs during testing phase

3.3.3 Study 1: Choice profiles

Binary tests

In the first experiment, only one monkey failed to meet the stringent preconditions for testing and the remainder showed a clear transitive choice bias an all remote pairs, AC, BD, CE, AD, BE and AE. The bias was highly significant on each pair ($p < 0.001$). The percentage of choices which had a transitive bias are shown in table 3–1. This table is in close accord with the child data (Bryant & Trabasso, 1971).

Acquisition

Figure 3–1 shows the relative percentage error rates (percentage of total error in a given phase) on each pair during the subjects original training. In phase one, training pairs were presented in 'runs' (as previously described) and in phase two they were in pseudo-random order. It can be seen that there appears to be no particular pattern to the errors in the first phase. Each pair contributes about 25% of the errors. During the random presentation, however, an 'inverted-U' shape emerges. This kind of curve is typical of memory studies in which subjects have to learn a list of unrelated items; retention is better towards the ends of the list. That subjects show the emergence of a strong invert-U curve as their performance improves, *before* encountering any remote pairs, is strong evidence against a model of inference in which pairs are simply 'stored' and then

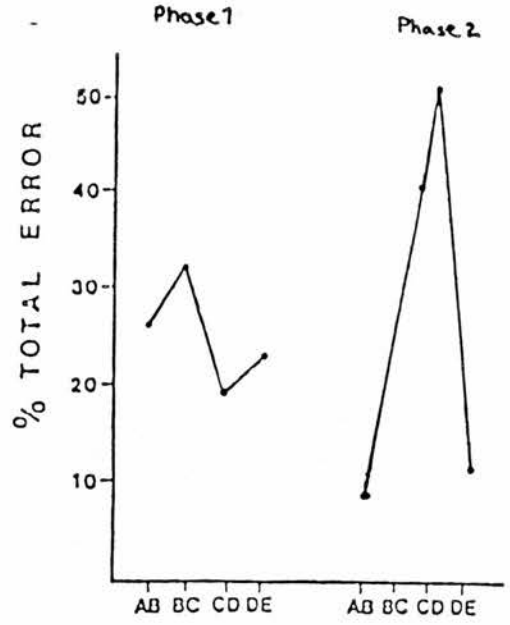


Figure 3–1: Acquisition profiles in monkeys

combined when the need arises. These curves are compared with equivalent results for children in a later section.

Triadic choice profiles (early)

Although performance was worse for the triads, it still showed a convincing transitive bias. Figure 3–2 compares the overall levels of transitive response on pairs and triads. The critical triad (containing no end terms) BCD is shown separately.

The binary sampling model

McGonigle and Chalmers proposed a 'non logical' account of how the monkey subjects performed the binary and triadic tests (McGonigle & Chalmers, 1977; McGonigle & Chalmers, 1984a). The idea is that triads and critical pairs are treated by subjects in much the same way, by sampling a binary subset of the relevant items and making a choice on the basis of this subset alone. For example given the triad BCD or the pair BD along with the 'inferred' item, C, there

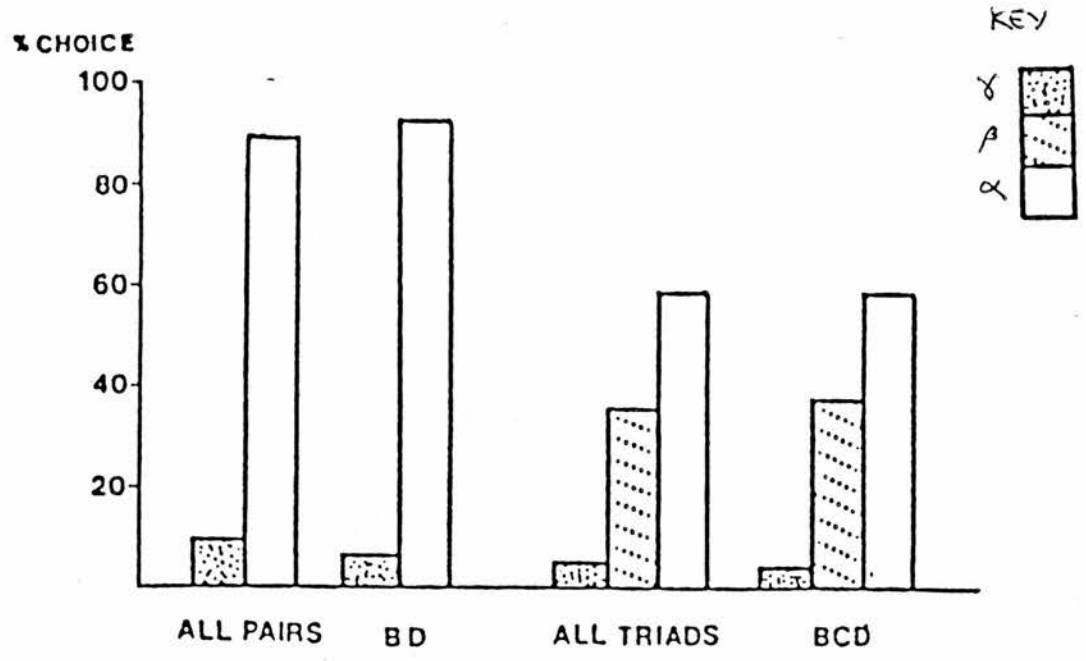


Figure 3–2: Distribution of choices to binary and triadic subsets of series

are three subsets which could be sampled, BC, CD or BD. It is assumed that subsets have an even chance of being sampled. Considering the triadic case first, what is the outcome of selecting each of the subsets? If BC is selected then C is chosen, as this is a training pair. Likewise, if CD is selected, D is chosen. Given BD, however, a random choice is made as there is no recourse to either training pairs or to uniquely labeled end-terms. Summing up the choices to individual items, this gives a one in three chance of selecting C, a one in two chance of selecting D (a third from CD plus a sixth from BD) and a one in six chance of selecting B. Over a hundred trials, therefore, we might expect the distribution of choices to be approximately 17% to B, 33% to C and 50% to D. In fact the actual performance was somewhat better than this, the proportions being 3%, 36% and 61%, respectively. Nevertheless, the overall fit of the projections of this model to the group data is quite striking, as reported in the original *Nature* article.

Table 3–3 shows how another example triad projection is made. It is assumed that if a sampled pair contains an end-term (A or E) or it is among the four training pairs, then the response is unambiguous. The implication is that, to

Triads	Choice projection*			Obtained		
ABC	00	33	67	00	31	69
BCD	17	33	50	03	36	61
BDE	17	17	67	16	24	60
CDE	00	33	67	11	24	65
BCE	00	33	67	06	28	66
ABD	00	50	50	00	44	56
ACD	00	33	67	00	30	70
ADE	00	33	67	01	21	78
ABE	00	33	67	00	30	70
ACE	00	33	67	00	26	74
Average distribution	0.03	0.33	0.64	0.04	0.29	0.67

*The figures in the left-hand column are predictions based on the assumptions that the subjects are sampled equally often and that preferences are absolute within those subsets presented during original training.

Projections of binary sampling model compared with *early* monkey triadic choice distributions. Adapted from (McGonigle & Chalmers, 1977).

Table 3–2: Binary sampling model and *early* monkey triads.

perform at the level they do on the triads, subjects need only have learned (a) the correct response to the training pairs and (b) the unique labeling of the end-terms as rewarded or non-rewarded. Table 3–2 compares the projections for all ten triads with the actual distribution of choices averaged across all eight subjects. It is interesting to note that even this 'non-logical' strategy allows for a graded response to each of the items such that there is a degree of seriation of the set of items. Table 3–4 shows the monkey choices displayed in another way such that the number of times each item is selected is summed. Each of the five items appears with equal frequency within the set of triads and so the apparent seriation can be compared with the null hypothesis that each item has the same chance of being selected. The projection of the binary sampling model has been included for a further comparison.

Difficulties with the binary sampling model

Although the binary sampling model is plausible for the triadic (group) data, it does not seem so attractive in the case of the pairwise comparisons. Going

Subset	*B*	*D*	*E*
BD	1	1	–
BE	0	–	2
DE	–	0	2
Totals	1	1	4
%	17	17	67

Table 3–3: Example showing how projection is made for a single triad

Triad	*A*	*B*	*C*	*D*	*E*
ABC	0	22	48	–	–
BCD	–	2	25	43	–
BDE	–	11	–	17	42
CDE	–	–	8	17	45
BCE	–	4	20	–	46
ABD	0	31	–	39	–
ACD	0	–	21	49	–
ADE	1	–	–	15	54
ABE	0	21	–	–	49
ACE	0	–	18	–	52
Totals	1	91	140	180	288
%	0	13	20	26	41
Sampling	0	15	22	23	40

Table 3–4: Overall triadic choice matrix for monkeys showing frequency of choice within triads and for individual items. Bottom rows show percentages of total number of choices and the percentages predicted according to the binary sampling model. Note the gradation of response according to serial position

back to the previous example, the model says that, given the pair BD, the item C must be inferred by common association with the other two items. The same process then occurs as for the triad BCD except that all choices which would have gone to C now go to D: '... the choice proportions attributable to C when actually present will now add to the overall proportions for D (as half of them will rule out responses to B, the other half will confirm D directly). The probability values for BD in a two-choice situation will now be: $B = 0.17$, $D = 0.83$.' (McGonigle & Chalmers, 1977). It is not at all clear how choices attributable to C could 'confirm D directly' because C is only chosen when the pair BC is sampled. The process would have to be one of rejecting B and thus picking D by default.

The account is vague on the issue of how other remote pairs are dealt with. For example, the comparison of A and D could elicit either or both of the 'reference' items B and C, making the decision process still more complex. This would appear to go against the *ordinal distance effect* finding (described for monkeys in a later section), whereby adjacent pairs are the slowest responded to and more distant pairs are faster. The problem of specifying how 'internal' items are 'inferred' is also more acute for the remote pairs. One possible way round these problems might appear to be for subjects to deal with pairs involving end terms firstly and separately, without bothering to infer intermediate terms and sample from them. A problem with this is that the model is then predicts subjects would be limited to five-term series. They would not be able to compare, say, items 2 and 5 in a six term series, whereas Trabasso *et al* have shown children can learn six-term series - there is no reason to suppose that monkeys could not. There is also recent evidence from a study of transitivity in pigeons (section 3.2) that the presence of an explicit dimension of difference between stimuli does not significantly affect performance.

There are still more reasons for rejecting such a compromise. First, the ordinal distance effect exists even amongst the 'internal' items, as we shall see below. Finally, if decisions are made first and foremost on the basis of the selection or rejection of end-anchors, then this undermines the rationale for sampling in the

triadic case. For example, given that the triad BDE then, to be consistent with the binary strategy, subjects should first look for an end-anchor, in this case, E. As this leads to a decision straight away (select E) then there is no point in sampling a binary subset.

3.3.4 Study 2: Choice Profiles

Binary choices

Following retraining, all five subjects showed a strong transitive choice profile, including one subject who performed weakly on the BD pair the first time round.

Middle Phase

This short testing phase showed a similar drop in performance, in comparison with the binary tests, as occurred in the first study.

Late Phase

Subjects showed an improved performance compared with the *middle* phase, despite having received no intervening training and no selective feedback as to whether their responses were correct or incorrect[3]. Figure 3–3 shows the change in performance, with the original phase also shown for comparison.

3.3.5 Study 2: Monkey reaction times

A major initial finding was that the monkeys showed an *ordinal distance effect.* Figure 3–4 shows the group effect for all pairs and separately for the 'internal'

---

[3]The subjects were explicitly trained to perform the triads after the *late* phase to see if their performance could be tuned up still further. With selective feedback, performance became almost perfect.

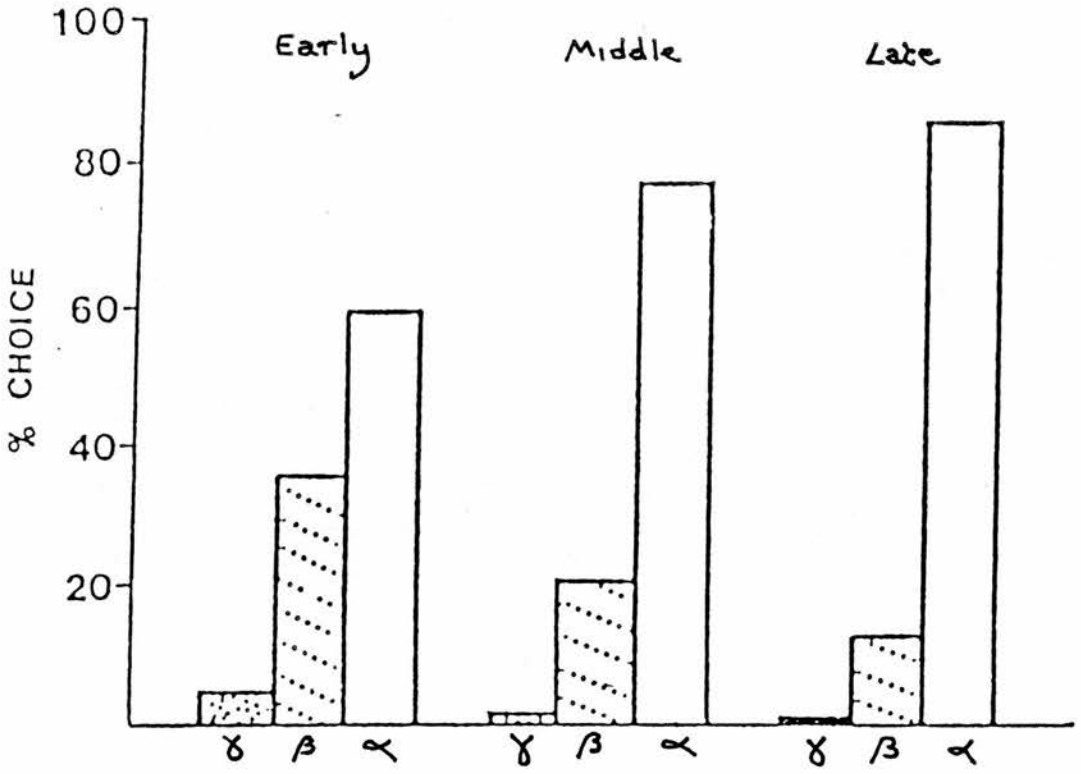


Figure 3–3: Improvement in triadic performance across testing phases

pairs. Figure 3–5 shows that this effect is not merely an emergent group phenomenon but exists for each individual (*albeit* with less linearity in some cases). This is the only study in which the distance effect had been demonstrated in 'non-logical' subjects, and the finding of the effect in individuals represents the most convincing evidence yet obtained that subjects are employing an integrated representation.

3.3.6 Comparing Children and Monkeys

McGonigle and Chalmers carried out analogous transitivity experiments with children. The experimental procedures employed were similar enough to those they employed with monkeys to enable a detailed cross-comparison of performance to be made (McGonigle & Chalmers, 1984a). In the first experiment, six year old children were given a non-verbal version of the five-term series task, using the same stimuli and basic procedure as employed with the monkeys (described above). Instead of peanuts, a counter was hidden under one of the tins in each training trial, and subjects were encouraged to collect these and exchange them for sweets (candy) at the end of the experiment. When children made

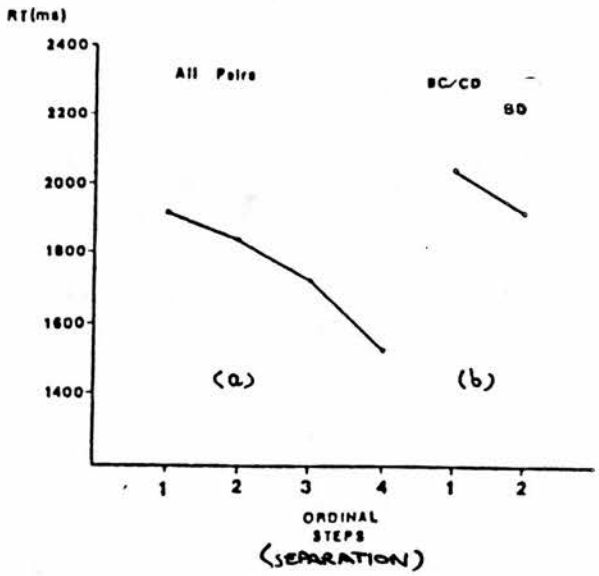


Figure 3–4: Ordinal distance effect in monkeys ($N = 5$) for all pairs (a) and for non-end-anchor pairs (b).

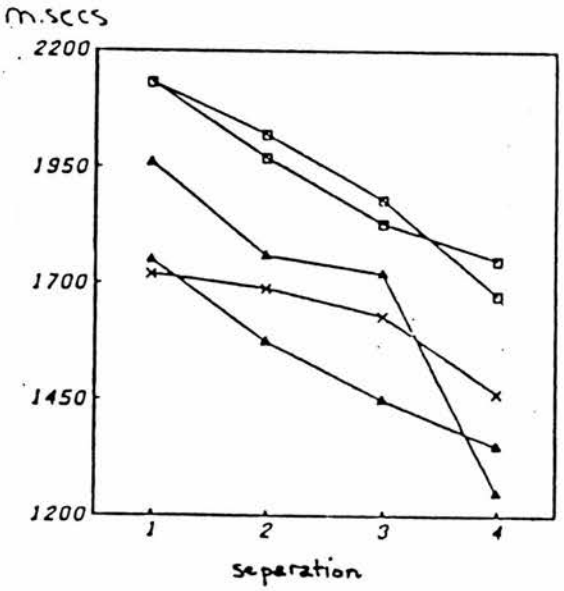


Figure 3–5: Monkey ordinal distance effect plotted for individual subjects

correct choices the experimenter said "That's right", when wrong, they were told "No, thats wrong, isn't it?", and the experimenter demonstrated that the counter was hiding under the other tin. The training schedules paralleled those of the monkeys (starting with the pairs presented serially and then switching to random presentation), although the childrens' training was shorter and less intensive. The acquisition profile is shown in table 3–6 below, and this can be compared to the corresponding monkey data (albeit with one less phase) shown in table 3–1. The three phases for the children are (1) *serial* (runs of AB, BC, CD, DE in order), (2) *recursive serial* (serial but with each pair presented four times consecutively) and (3) *random* (trials were presented in randomly ordered blocks of the four different pairs). It can be seen that children and monkeys' learning curves appear similar, with both tending towards an inverted-U shape. The only major difference is that the children appear to start of by showing a strong 'recency effect' in which the most recently presented pair is remembered better. It is not known why they should show this and not the monkeys, but it could be an artifact of differences in experimental procedure during early training. Certainly the most salient aspect of the two experiments, taken together is the way both sets of subjects 'grow' into a pronounced serial position curve.

Binary tests

Table 3–5 below, shows the results of the subsequent testing phase compared with the monkey results. As with the monkeys, the six remote pairs were tested without differential feedback, although corrective feedback was maintained on the adjacent pairs. All test responses were significantly biased ($p \leq 0.01$) except for the pair AD. It seems likely that the lack of a significant bias on this pair is due to the relatively short training that the children received and the small sample size. Inspection of the data from individuals shows that most subjects showed a stronger bias than emerges for the group and that two or three subjects showed a bias in the reverse direction on the crucial pair BD. Unfortunately, there is not enough data to support separate statistical tests for individuals. It can be seen that there are some similarities between the two tables but it is

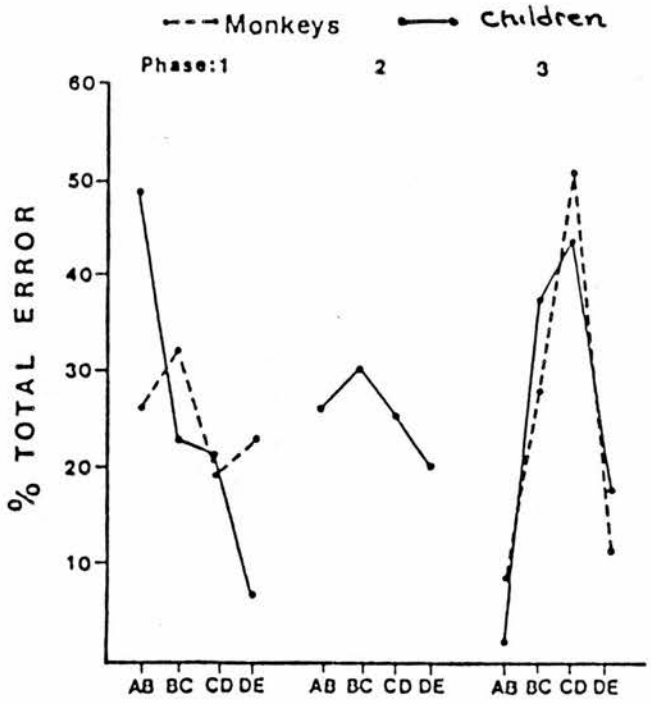


Figure 3–6: Acquisition profiles for six year old children compared with monkeys

difficult to compare the patterns because of the ceiling effect for the monkeys. It seems that, by casual inspection, the tables are almost mirror images, with high performance on the E pairs for children and high performance of the A pairs for monkeys.

	B	C	D	E
A	98	70	57*	80
B	–	98	70	88
C	–	–	98	78
D	–	–	–	98

Children (N=10)
*Non significant on a binomial test

	B	C	D	E
A	98	100	100	100
B	–	93	90	76
C	–	–	89	87
D	–	–	–	97

Monkeys (N=8)

Table 3–5: Child and monkey binary choice profiles

	Children			Monkeys			Bin.Samp.		
Triad	γ	β	α	γ	β	α	γ	β	α
ABC	8	28	64	0	31	69	0	33	67
BCD	20	23	57	3	36	61	17	33	50
CDE	17	15	68	11	24	65	0	33	67
ABD	10	43	47	0	44	56	0	50	50
BCE	3	47	50	6	28	66	0	33	67
BDE	28	20	52	16	24	60	17	17	66
Means	14	29	57	6	31	63	6	33	61

Child and monkey choice profiles compared with binary sampling model projections (only six of the possible triads are shown). Data from (McGonigle & Chalmers, 1984a).

Table 3–6: Child and monkey triadic test results

Triadic tests

Table 3–6 shows the children's triadic test profiles compared with the monkeys'. The children were only given six of the ten possible triads (to avoid making the tests too laborious for the subjects). It can be seen that children, like monkeys, show a marked reduction in transitive bias compared with the binary tests and both show a reasonable concordance with the projections of the binary sampling model. Unfortunately a statistical comparison between choice levels in the two species is not possible as monkeys were given more presentations of each pair and triad as well as a larger set of triads (all ten) than were given to the children.

Verbal Training

Despite the overall similarity between children and monkeys on the above non-verbal version of the five-term series task (as indicated by the acquisition profiles, successful binary test performance and the reduction in transitive bias on the triads), it could be argued that the training procedure somehow induces an atypical 'non-logical' strategy. In particular, the training is non-verbal and effectively

involves only one comparative for any given subject (*eg* choose heavier). This is in contrast to (Riley & Trabasso, 1974), who found that linguistic training using both complements of a comparative (A is longer than B and B is shorter than A) produced better retention on the pairs BC and CD. These authors concluded that use of only a single comparative promoted use of nominal rather than ordinal encoding strategies, and thus gave rise to poorer performance.

To help settle this issue, McGonigle and Chalmers conducted a second experiment (published in the same paper) in which the only difference in procedure was that the children had to answer verbal questions instead of just selecting the 'correct' tin. In both training and testing subjects had to respond to each trial by answering one of the questions "Which is the heavy one?" or "Which is the light one?". Question form varied randomly from trial to trial and feedback (during training) was verbal, *eg* "That's right, green is the heavy one" or "No, that's wrong, green is the heavy one". A 'reward' (sweets) was given at the end of every training session to encourage participation.

As the structure of this experiment was the same as the previous one (in terms of length of training and testing *etc*), a direct statistical comparison of verbal and non-verbal performance was possible. A two way analysis of variance with repeated measures revealed no significant main effects or interactions between the verbal/non-verbal grouping and test pair. Triadic performance was also highly concordant (14; 29; 57% and (14; 34 ;52%) to γ, β and α items for non-verbal and verbal groups.

In addition to the binary and triadic tests, subjects were also presented with a verbal seriation task. With no stimulus present, the child was asked, "Which is the heaviest tin?", "Which is the next heaviest", and so on until they gave up or repeated earlier colours. As an additional test for ordinal encoding, subjects were shown one tin at a time (in random order) and asked to categorise it as heavy or light purely on the basis of seeing its colour. Both these post-tests showed that subjects could only locate the ends of the series with any reliability. The results of the verbal seriation task are shown in table 3–7. This apparent inability to seriate, combined with the reduced performance on the triads, is

Item	*A*	*B*	*C*	*D*	*E*
p	37*	15	29	26	37*

Probability of assigning each item its correct position in the series. *Significantly deviant from chance (20%). Data from (McGonigle & Chalmers, 1984a).

Table 3–7: Verbal seriation post-test (N=19)

evidence against the idea that an ability to mentally seriate objects underpins performance in transitive inference tasks.

3.4 Conclusions

1. It appears that the ability of subjects to show a transitive bias on the crucial *BD* pair in the five-term series task is a phenomenon which is robust with respect to species, methodology and stimuli.

2. The ordinal distance effect, 'reduced' performance on triadic tests and an inverted-U shaped serial position curve during acquisition all appear to be indexical of subjects strategies.

3. 'Symbolic' differences between the stimuli determine the representation and that it is not crucial to have the presence of a physical metric such as size.

4. Subjects appear to employ a 'non-logical' strategy even in verbal versions of the five-term series task.

5. Ability to seriate (assign each object a unique location within the series) does not appear to be a prerequisite for solving the five-term series task.

6. Given the strong concordance between the child and monkey data, it is reasonable to assume that, in this domain, monkeys' strategies are representative of those used by children.

Clearly, it must be accepted that subjects are not doing 'formal' (deductive) transitive inference in these tasks. Breslow is right in this respect. Monkeys, or young children, cannot be assumed to be logical symbol-manipulators by birthright, and it is unreasonable to suppose that they can acquire a formal theory of transitivity during the course of the five-term series task. Such skills need to be acquired through teaching, after all, it took two thousand years for modern logical calculus to be developed from its roots among the Greek philosophers. Furthermore, transitivity is not inherently 'in' the environment and so cannot be acquired, in the way Piaget suggests, by internalising actions. The available evidence suggests that rather than the ability to physically seriate underpinning the emergence of formal transitive reasoning, its more the other way round; a 'pre-logical' form of reasoning involving linear orders is the primitive on which these other skills are built. This primitive form of reasoning, which appears to be captured in the five-term series task, can still be regarded as *inference*, in the psychological sense discussed in the previous chapter. These issues are returned to in the concluding chapters.

As the monkey studies are the only ones to have a rich data base for each individual subject (thus supporting statistical analysis for individuals), the monkey data provides the only viable source for a microanalysis, based on computational modelling, of subjects' strategies during the five-term series task.

Chapter 4

The Stack Model

This chapter introduces and specifies a simple rule-based model of the decision mechanism underlying subjects' choices during performance of the five-term series task. It is shown that the model can account qualitatively for the main (group) phenomena associated with the child and animal data introduced in the previous chapter. Some possible variations on the model are considered but a simple core model is put forward as the best candidate for assessment.

4.1 A task Analysis

In order to produce a computational model of a typical subject's strategy, it is first necessary to form a "monkey's eye" impression of the five-term series task (described in the previous chapter). It is difficult to do this for competent performance alone — a task analysis almost inevitably raises questions about what knowledge subjects bring to the task *vs* what they learn from the task.

What prior knowledge can we safely assume the subjects possess? Firstly, as the animals had some prior testing experience, it seems reasonable to assume that they had some knowledge of the generic form of the task which faced them. They would know they had to choose one from a (small) number of differing stimuli and that (at least?) one of the possible choices would be rewarded and, possibly, that one would be unrewarded. Children performing the task might be told as much directly.

The nature of other prior knowledge has to be inferred more indirectly. From the subject's point of view, rather than having to 'make an inference' (a request which would be difficult to communicate to children, let alone animals), the task must surely be to learn how to use the characteristics of the stimuli to predict which future choices will be rewarded. From this perspective, it is not safe to assume that subjects will either possess or employ a transitive inference schema; or that they will organise the knowledge they pick up, in any particular format. We may assume that subjects will bring to bear *some* general cognitive apparatus for dealing with situations of this type, although we cannot say *a priori* what this will be. There may also be special purpose mechanisms brought to bear but, again, we cannot say what these are without further evidence.

Let us now consider the task itself. The training stimuli which the subjects had to make sense of are shown schematically below. Letters represent colours, the location of stimuli is represented by their left-right relation on the page. A '+' underneath a letter indicates a (hidden) reward associated with the corresponding object and '0' indicates no reward.

A	*B*	*B*	*C*	*C*	*D*	*D*	*E*
0	+	0	+	0	+	0	+
B	*A*	*C*	*B*	*D*	*C*	*E*	*D*
+	0	+	0	+	0	+	0

It can be seen that there are eight basic training situations which could potentially be discriminated prior to each choice. However, it must be remembered that this is an abstraction; there are multiple occurrences of each of the situations in an unpredictable order [1]. From the subject's point of view, there is only a succession of trials which must each be dealt with; the subject cannot know in advance (or ever be certain) how the set is bounded. At any stage a novel

[1]Not completely random — early trials were presented in a systematic order.

situation might arise which would break previous rules or categories. Indeed, this is just what happens when the testing phase starts.

Given this state of affairs, what are the space of possible inductions? In other words what generalisations can be made about the above training instances? Here are some examples of possible generalisations, expressed as rules, using a production systems shorthand: $< condition > \Rightarrow < action >$.

1. $left(C)$ & $right(B) \Rightarrow rewarded(C)$.

2. $present(B)$ & $present(C) \Rightarrow rewarded(C)$.

3. $present(E) \Rightarrow rewarded(E)$

4. $\neg rewarded(A)$

5. $present(D)$ & $absent(E) \Rightarrow rewarded(D)$.

6. $\exists x \quad present(x)$ & $\neg rewarded(x)$
   (At least one of the choice objects in a trial is not rewarded.)

7. The colours form a series, A, B, C, D, E and from any subset of these colours the item nearest the E end will be rewarded.

8. $present(C)$ & $absent(D) \Rightarrow rewarded(C)$.

The first example may, at first sight, appear not to be a generalisation at all because it only covers one case. However, this case is itself a generalisation, as pointed out above. The second example abstracts out location as being unimportant. It employs the higher level concept of presence of an item within a particular trial. The next few rules capture the same kinds of principles as used in the stack model. The seventh example is the most abstract and might seem the least likely to arise from a general inductive process, yet this is the kind of thing that is generally supposed is learned in the transitivity task. Up to this point all the generalisations make 'correct' predictions for the remote pairs

(where they are applicable). The last example is included to show that this is not necessarily the case. This rule is compatible with all the training instances and yet leads to a 'non-transitive' prediction for the pair CE.

Clearly such inductions are potentially useful. The question of how the inductive process itself might work is postponed until chapter 7. Our main concern for the moment is to decide what abstractions are employed by a competent performer of the task. The model which is described below was developed according to 'Occam's razor' principle that the number of objects within a theory should be minimised. It was decided to employ a production systems notation for the reasons outlined in chapter 2. Since the transitive reasoning task is essentially a symbolic inference problem, it was felt that it was likely that at least some aspects of the information gain processes could be captured using a sequential rule based system.

4.2 A model of binary performance

In outline, the model is based around the concept of *avoidance* and *selection* of objects along with a simple control element. A strategy for performing the task consists of having a small set of rules, each one of which is an instruction to either avoid or select a particular feature, along with a control element which attempts to apply each rule in turn in a fixed order. The strategy lends itself to a simple production system notation of a set of conditional rules, as shown below. The rules are applied to the task by an *interpreter* or *control structure* which, in this case, basically tries each rule in turn until one succeeds. For example, the series A, B, C, D, E can be represented by the stack of four rules and interpreter shown below:

STACK

1) $present(E) \Rightarrow select(E)$
2) $present(D) \Rightarrow select(D)$
3) $present(C) \Rightarrow select(C)$
4) $present(B) \Rightarrow select(B)$

INTERPRETER

1) Remove a rule from the top of the stack.

2) If the rule is not applicable to the trial then go back to step 1.

3) Carry out the action of the rule and stop.

In order to make the critical BD comparison, the interpreter would scan down the stack to the first relevant rule, which in this case is the third. The action of the rule leads to the choice of D This example rule stack also gives correct responses to all the other pairs. For example, all the pairs containing E are discriminated by the first rule, whereas the interpreter only reaches the fourth rule for one pair, BC.

4.2.1 Variants of the stack

There are a number of variants of the example stack which will also give correct performance. For example, the second and third rules can be swapped around. However, the third and fourth rules cannot be swapped around without giving an incorrect choice on the pair CD. A second source of variation is in the rules themselves, which can each take one of two forms. This particular rule set consists of three 'selection' rules and an 'avoidance' rule (no. 2). Either form of rule can refer to any one (and only one) of the objects in the series. For example, the last rule could be 'if B is present then avoid B' and the stack will still give correct answers. In fact there are sixteen (2^4) stacks of this generic form which will 'perform' correctly on all pairs from the five-term series. These range from a stack consisting of all selection rules through various mixed stacks such as the previous example to one consisting entirely of avoidance rules. The range is illustrated in table 4–1. A six-term series would require five rules per stack and there would be $2^5 = 32$ variants.

Note also that the order of mention of the items down the stack does not necessarily correspond to the order implied by the *performance* (*ie* the series A to E). This is in contradistinction to previous linear representation models as discussed in chapter 6.

1) $present(E) \Rightarrow select(E)$

2) $present(D) \Rightarrow select(D)$

3) $present(C) \Rightarrow select(C)$

4) $present(B) \Rightarrow select(B)$

$\ldots$

1) $present(A) \Rightarrow avoid(A)$

2) $present(E) \Rightarrow select(E)$

3) $present(B) \Rightarrow avoid(B)$

4) $present(D) \Rightarrow select(D)$

$\ldots$

1) $present(A) \Rightarrow avoid(A)$

2) $present(B) \Rightarrow avoid(B)$

3) $present(C) \Rightarrow avoid(C)$

4) $present(D) \Rightarrow avoid(D)$

Table 4–1: The range of rule stacks for solving the five-term series

Although there are as many as sixteen possible rule stacks which satisfy the constraint of performing correctly on all the pairs, the space of stacks which do not satisfy this constraint is even larger. Given five items and two forms of rule and then the number of stacks which contain four different rules is $10 \times 9 \times 8 \times 7 = 5040$ stacks. It is interesting to note, therefore, that a property of this space is that *any stack which performs correctly on the adjacent pairs also performs correctly on the remote pairs.* This is important because subjects only receive training on the adjacent pairs. Our model is therefore that subjects construct a stack for the purpose of dealing with the training information, and that this then (somewhat indirectly) gives them the ability to deal with remote pairs. The process by which stacks might be constructed is discussed in chapter 7.

4.3 The Full Stack Model

The basic control structure described above does not specify what happens if the task is to select between more than two items. For example, given the triad ACD then the following process results (with the same stack as before): Rule 1 does not match any item so the next rule is taken off the stack: 'If A is present then avoid A'. This eliminates A as a candidate but leaves the subject with no way of making a choice between C and D. Other triads, such as BCD present no such problems, with rule three making an appropriate unique choice in this case. One possible strategy would be to simply make a random choice where a single application of a rule fails to give a unique choice. This strategy is called *naive* and an interpreter is shown in figure 4–2.

The *naive* strategy makes a unique choice for each triad, but errors will be made where *avoidance* rules are involved. This turns out to be a good model of the *early* triadic tests, as shown in chapter 5. An implication of the *naive* control strategy being employed is the existence of two basic types of triad (with respect to a given rule stack). The first type, henceforth called *random* triads, have an expected distribution of 50% of choices to each of the α[2] and β items with zero choices to the γ item. *Random* triads are those where the shallowest applicable rule is an *avoidance* one. The second type, *alpha* triads, are those where the shallowest applicable rule is a *selection* one, because it expected that all the available choices will go to the α item.

A final point about the *naive* strategy is that the fourth rule never gets reached when the stack is applied to a triad — for any of the ten possible triads, one of the first three rules is always applicable, whatever stack is employed. This is simply because three rules are relevant to three out of the five items from the series, and a triad cannot be constructed from the remaining two.

---

[2]These symbols are defined in chapter 3.

4.3.1 Variants of the interpreter

In the cases like *ACD* (with stack 3), where an avoidance rule is the shallowest applicable rule, a more sophisticated control regime is needed which allows more than a single rule application. For example, the second control structure in table 4–2 makes sure that deeper rules are considered after the shallowest one has fired. Such a strategy is referred to as *iterative*.

These two types of control strategy (*naive* and *iterative*) show how the control strategy affects the adequacy of performance. In particular, a strategy which gives correct performance on the binary task will not necessarily perform correctly on triads. Performance on the binary tests is more 'robust' in the sense that the control structure is not so critical.

Partial iteration

Another aspect of the triadic testing phase is the 'spontaneous' improvement found in the second study. This can be accommodated into the above framework by assuming that the new tests lead the subject to develop a more principled control structure for the task. In other words they generalise their strategy so that it is not limited to pairs but can make a unique choice between a number of items. This seems plausible, as it is only on the triadic tests that the shortcomings of a simple control structure become apparent. However, subjects do not appear to make a sudden jump in performance as they switch strategies but rather show a more gradual improvement over the three phases. It seems that there is a need to hypothesise some intermediate kind of control strategy, or a process of transition.

The only difference between the *naive* and *iterative* strategies (as specified) is in (the last clause of) the fourth step. If this rule does not suddenly get changed, there could be a transition period during which both forms of step 4 are kept around (pending evaluation of the new form). This idea is encapsulated in the third strategy in table 4–2, named *semi-iterative*. If, upon reaching step 4, there was (say) a 50% chance of either version being used then this would lead to

Example Stack:

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

NAIVE

1) Remove a rule from the top of the stack.
2) If the rule is not applicable to the trial then go back to step 1.
3) Carry out the action of the rule.
4) If unique choice then stop, else make a random selection.

ITERATIVE

1) Remove a rule from the top of the stack.
2) If the rule is not applicable to the trial then go back to step 1.
3) Carry out the action of the rule.
4) If unique choice then stop, else go back to step 1.

SEMI-ITERATIVE

1) Remove a rule from the top of the stack.
2) If the rule is not applicable to the trial then go back to step 1.
3) Carry out the action of the rule.
4) If unique choice then stop, else —

   — either go back to step 1 or make a random selection.

Table 4–2: Three control strategies for the stack model together with sample rule stack

an intermediate level of performance of the triads. Returning to the example rule set and the triad ACD, rule 2 is applied (as before) at step three in the strategy, thus eliminating A. At step three the process will either terminate with a random choice between C and D or it will recurse back to step one. Over a number of trials on which the former happens, the average distribution of choices between A, C and D will be 0%, 50% and 50%, respectively. In the latter case, rule 3 will be invoked at step 3, and the process will terminate at step four with the unique selection of D (and a resulting distribution of 0%, 0%, 100%). The net result of applying the *semi-iterative* strategy would thus be a distribution of 0%, 25% and 75%, halfway between the other two strategies.

Naturally, a different probability weighting between the two parts of step four would lead to a proportionally different outcome. In the next chapter this variable factor is referred to as the *percentage of iteration* (the probability of carrying out the first part of rule 4). Over a number of decisions, this is equivalent to the proportion of trials on which more than one rule is applied. Note that the *naive* and *iterative* strategies can be viewed as special cases of the *semi-iterative* strategy, with 0% and 100% iteration, respectively. Also, of course, any strategy will work for the binary pairs where avoidance rules are unambiguous.

This completes the description of the stack model. It was considered that three control strategies (combined with the sixteen possible rule stacks) give enough degrees of freedom to make an interesting modelling attempt. The final section in this chapter contains a discussion of further possible control strategies which, it was considered, would make the evaluation process (in the next chapter) too complex.

4.4 Summary of the Stack Model

The basic model to be evaluated with respect to the monkey five-term series data is summarised below.

1. There are four rules per subject which are relevant to a set of five objects: A, B, C, D and E. Alternatively, these may be thought of as five distinguishing properties (for example, colour) of a set of objects.

2. Each conditional rule may have two forms, *selection* and *avoidance*:

   - selection: $present(X) \Rightarrow select(X)$
   - avoidance: $present(X) \Rightarrow avoid(X)$

   Where X is one of the five items.

3. The four rules allow the correct choice to be made on all training pairs (and hence on all the pairs).

4. The forms and order of the rules are invariant for a given stack.

5. The stack is interpreted by one of three control strategies (to be specified in the evaluation): *naive*, *iterative* or *semi-iterative* as outlined in table 4–2.

6. The *semi-iterative* control strategy has a variable component which specifies the degree of iteration. This can vary from 0% (equivalent to *naive* to 100% (equivalent to *iterative*).

Variables

1. Rule form and order can vary between subjects.

Identity	Rule depth			
	1	2	3	4
*1.	Se(E)	Se(D)	Se(C)	-
2.	Se(E)	Se(D)	Av(A)	-
3.	Se(E)	Av(A)	Se(D)	-
4.	Se(E)	Av(A)	Av(B)	-
5.	Av(A)	Se(E)	Se(D)	-
6.	Av(A)	Se(E)	Av(B)	-
7.	Av(A)	Av(B)	Se(E)	-
*8.	Av(A)	Av(B)	Av(C)	-

Table 4–3: Enumeration of stacks for modelling purposes

2. In the case of the *semi-iterative* control strategy, the relative probability of making a random choice or recursing can vary between subjects or for the same subject across time.

Enumerating the stacks

There are sixteen possible rule sets which satisfy the above constraints. For shorthand purposes, it will be useful to enumerate these explicitly. The numbering system that has been adopted here is geared to the purposes of evaluation. For each possible sequence of rule forms, there is one corresponding set of rules which give correct performance on all the pairs. The ordering is based on a binary sequence progressing form all *selection* to all *avoidance*. However, we need only consider eight stack forms, and these are enumerated in table 4–3, with the form of the fourth rule unspecified (rule forms are denoted by their first two letters). The stack used in the examples corresponds to number 3 in this schema.

The form of the fourth rule has few implications for the behaviour of the rule stack. This is because the last rule is only ever used to discriminate one pair, the other possible binary (and triadic) combinations being dealt with by rules

higher in the stack. For evaluation purposes, it makes little difference to the overall projected performance profile whether the last pair is discriminated on the basis of selection or avoidance. The simplifying assumption will be made, therefore, that the fourth rule is always a selection one, unless otherwise stated. For the sake of completeness, the remaining eight rule stacks (with an avoidance rule at the bottom) may be referred to as 1(av), 2(av) *etc.*

4.5 Implications and global fit of the model

This section describes the ways in which the stack model demonstrates the typical group phenomena (described in the previous chapter). Also discussed, are the implications of the model for evaluation at a more detailed level.

4.5.1 The ordinal distance effect

If the main source of RT variation is the depth of the rule which 'fires' then **any set of rules fitting the specification will show an ordinal distance effect.** This is assuming that the length of time it takes to carry out selection and avoidance actions are not too dissimilar[3]. An example of how an RT profile can be generated from a set of rules is given below.

Assuming that it takes one unit of time to pull a rule off the stack, one unit of time to test a condition and two units of time to carry out the action and displace the object (the assignments are not critical), then the predicted RTs for each pair can be tabulated along with the mean RTs for every ordinal separation (table 4–4). The mean RTs along the diagonals of this table give the projected

[3]The difference must not be larger than the time to fail a condition and move onto the next rule. This works out to be equivalent to assuming that the main source of variation in reaction time is the number of rules that are tried before one succeeds, or the depth of the successful rule.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

	B	C	D	E
A	6	6	6	4
B		10	8	4
C			8	4
D				4

ordinal separation	mean RT
1	7
2	6
3	5
4	4

Table 4–4: Projected RTs show distance effect for stack No. 3

distance effect as shown alongside. Any of the other stacks will give the same distance effect — the tabulated RTs look similar except for permutation of the entries along the diagonals (the AE comparison always takes 4 units, as the top rule in every stack must be either about A or E).

Note that the same rank differences in the RTs are produced by taking the depth of the rule employed to discriminate each pair as a measure. An alternative explanation for the reason why the distance effect emerges for any stack is that, the further apart two items are in the series, the more likely it is that one of them will be 'caught' by a rule high up in the stack.

Linearity hypothesis

Although the model is thus in accord with the monkey phenomena in showing a distance effect at both the group and individual levels, it also predicts a more fundamental effect at the individual level. That is, it should be possible to plot a linear relationship between an individual's reaction times and the depth of rule

employed to make a decision. In the actual data, this will be referred to as the *depth effect*. The prediction can be broken down into two parts:

Grouping hypothesis: Pairs involving the same decision process should produce the same RT. For example, given the stack in the previous example, the pairs AB, AC and AD should all have the same RT.

Linearity hypothesis: The variable component of RT should be proportional to the depth of the rule that provides the choice. A weaker version of this hypothesis states that the RT should monotonically increase with rule depth.

The rank ordering of reaction times from fastest to slowest can thus be represented as follows, for stack No. 3:

1) *AE BE CE AE*

2) *AB AC AD*

3) *BD CD*

4) *BC*

4.5.2 Acquisition curves

The stack model as presented makes no prediction about error rates on the binary tests except for the possibilities mentioned in the discussion of resource limited strategies (section 4.6.2). If there were errors correlated with the depth of search of the stack then this would lead to an 'inverted-U' curve for the error rates on adjacent pairs when averaged across a group of subjects using different stacks. The serial position curves for individuals, however, could be asymmetrical or zig-zag in shape depending on the particular rule stack. Chapter 7 discusses how the curve could emerge as part of the learning process.

4.5.3 Triadic choice profiles

Section 4.3.1 describes how the stack model with the *naive* control strategy can account for the diminished levels of performance on transferring from the binary to the triadic tests. The *semi-iterative* control strategy can also model the increase in performance on the triads across time. Moreover, the model predicts that individuals will show differing characteristic patterns of selection and avoidance on the triads, depending on which stack they are employing. This fact will be exploited in evaluating the model in the next chapter.

For example, stack No. 1 will give correct performance on all the triads (irrespective of control strategy). This is because it contains only selection rules which always specify a unique choice. At the other extreme, stack number 8 will give a poor performance on all the triads. Versions of stacks 1 and 8 are shown in table 4–1. Intermediate stacks lead to a mixed performance whereby perfect performance is attained for some triads and a diminished performance for the remainder. The factor determining performance for any particular combination of stack and triad is whether the shallowest relevant rule in the stack is selection or avoidance.

4.5.4 Conclusions

At a gross descriptive level, the stack model appears to be able to account for most of the phenomena associated with the binary tests and for the drop in performance on transferring to the triadic tests. This is an encouraging start but begs the question of whether the model would stand up to a quantitative comparison with the data, particularly the triadic choice patterns.

It can now be seen why the monkey studies are the only appropriate source of data for such an analysis. This is partly because of the peculiar nature of the stack model in that, unlike its predecessors, it claims to be able to model individual subjects better than group data. Only the monkey research has collected enough data on individuals to make this feasible. A second reason for preferring the data from this study is the triadic post tests given to subjects after they had

learned the series. If we are correct in our assumption that subjects continue to use the same basic representation on transferring to the triads, then these provide a unique 'second window' onto the underlying processes, and a critical test for any model.

4.6 Variations on the stack model

The following variations on the stack model are not explicitly evaluated in the following chapter. They are included mainly for discussion purposes, but may be referred to with respect to potential solutions where the stack model breaks down.

4.6.1 Dynamic ordering of rules

A major assumption with the stack model is that subjects use the same rule stack throughout binary and triadic testing. Whilst the possibility of dynamic rule ordering is considered for the acquisition phase of the binary task (chapter 7) it is assumed that subjects would stick with the same set of rules once success was established. The possibility that subjects might re-order rules to optimise their performance in the triadic phase should not be ruled out. Similarly, rules might be added to or dropped from the stack.

Instead of a stack-like rule ordering in which each rule has an unambiguous (discrete) position, rules could have different (numerical) *priorities* and the control strategy could operate with respect to these. If each rule has one of four discrete, ascending values attached to it then the predictions of this version of the model remain much as before. However, this mechanism allows the possibility of two or more rules having the same priority value, or of some values being closer to each other than others. This potentially allows a different kind of indeterminacy from that in the *random* type triads, previously considered. With such a mechanism, it is easy to imagine that the order of rules might not

be totally rigid but could fluctuate across time, depending on how stable the priority values were. Such a fluctuation might also, conceivably, be an integral part of the mechanism for acquiring a rule stack.

Assuming that (effective) rule order is not entirely stable, the high performance that most subjects achieve on the binary tests (previous chapter) would appear to constrain most of the possible fluctuations to those that do not affect binary performance. Adjacent *avoidance* and selection rules can be swapped around without causing problems on the binary tests. However, such swapping does have implications for the triadic tests, causing variation in the distribution of choices between α and β items, as should become clear in the next chapter.

4.6.2 Resource limited control

The idea behind this is that the improvement on the triads is correlated with some kind of increased allocation of computational resources the task. Correspondingly, the initial reduction in performance that subjects show when transferring from the binary to the triadic task is because the triadic tests are more computationally intensive. A control strategy which produces this kind of behaviour is partially represented in table 4–5. Each time the strategy is called upon to make a decision it is allocated a finite amount of computational resources (*eg* memory, processing time or their biological equivalents). This seems a sensible precaution in a real-time system; in general, if a process uses more resources than expected, the chances are that something has gone wrong. As the process runs, it monitors its use of resources, which will be depleted by its own actions. For our purposes, we will consider such imposed resource limitations as outweighing any absolute (hardware) limitations.

In the example, the monitoring process occurs in step 2, which acts as a kind of 'fail-safe' condition, ensuring that the process will not grind to a halt without some kind of decision being made. If the process was given unlimited resources, however, this condition would never arise and the strategy could make a unique selection between any subset of an arbitrarily long series (given an appropriate

1) If unique choice then stop.

2) If low resources then make random choice and stop.

3) Remove a rule from the top of the stack.

4) If the rule is not applicable to the trial then go to step 2.

5) Carry out the action of the rule and go back to step 1.

Table 4–5: Resource-limited control strategy

stack of rules). With limited resources (diminishing at each step of the process), however, the process may terminate before the choice has been narrowed down to a unique item.

Overall, this leads to very similar behaviour to the *semi-iterative* control strategy, but there are some subtle differences. For example, using this strategy the previous example example triad (ACD) and rule stack (table 4–2), the results are pretty much the same. Given resources which are only sufficient for binary pairs, A will be eliminated and a random choice made between C and D. Beyond this point, the more resources there are available, the deeper the stack can be searched. A slight increase is all that is needed in this example for the third rule in the stack to be reached ($present(D) \Rightarrow select(D)$) thus enabling a unique choice. However, this would not be sufficient to make a unique choice from the triad ABC, where the fourth rule needs to be reached.

In summary, all that is needed is some trial to trial variation (noise) in the resource allocation to this strategy and it gives rise to a similar set of choice distributions to the *semi-iterative* strategy. The difference is that there would be a gradation of performance between the triads depending on how deep the rules needed to be searched to make a unique choice. The advantages of this version of the model are twofold. Firstly, it can potentially account for a wide range of behaviour with the changing of a single variable (or, more realistically, set of variables) — resource allocation. Secondly, it gives a neat way of accounting for the noise in the choice profiles in both versions of the task. This could be

viewed as either due to noise in the levels of resources allocation itself or in the quantities of resources used during each step of the process itself. For example, if time is one of the resources, then distractions during the decision-making process might lead to the occasional random choice even in the binary tests. The main disadvantage is that the model is very difficult to evaluate precisely because it is powerful. A major problem would be establishing the relative costs of the diferent components of resource usage.

4.6.3 Item-driven control

The class of control structures which have been alluded to so far are all what might be called 'stack-driven', in that the order in which rules are retrieved and tested is governed by the stack order. This means that, where more than one rule is applicable to a given trial, the shallowest one will always be applied first. At the other extreme, rule selection could be entirely item driven, so that instead of retrieving a rule and testing its condition against the items in the trial; an item is selected, and then a rule found to fit it. The way the decision process works in this case is that firstly, all of the rules relevant to a particular trial are retrieved and secondly, they are applied in the order dictated by the stack ranking. This would appear to predict perfect performance on the triadic tests, however unless there was something analogous to resource limitation, in which case the predictions are similar.

Chapter 5

Evaluating The Stack Model

Having shown how the stack model is capable of accounting qualitatively for the group *phenomena associated with the monkey data, this chapter deals with the evaluation of the model more quantitatively. The approach is to assess the 'fit' at coarse levels of description and then to progressively chunk the data more finely until the inadequacies of the model become apparent. The analysis begins with triadic error patterns and goes on to look at reaction times during binary tests. The notation used to refer to the data differs slightly from that employed by McGonigle and Chalmers — see section 3.3.2.*

5.1 Methodology

It was decided that the triadic choice profiles, particularly from the *early* phase, should form a key role in evaluating the stack model. This is because the stack model makes clear quantitative predictions in this area which are different from its only rival in this area, the binary sampling model (see chapter 3). Another reason for putting the emphasis on the triadic data is that most subjects made very few errors in the binary tests, creating a 'ceiling effect' whereby the error rates are uninterpretable. Table 5–1 shows the distribution of choices amongst the remote pairs for each of the seven subjects (adjacent pairs were not tested without differential feedback). It can be seen that whilst the first five subjects show a near perfect performance, White is a little erratic on two pairs and Green actually has a reverse bias on one pair.

⋱	*A*	*E*	*A*	*C*	*A*	*D*	*C*	*E*	*B*	*E*	*B*	*D*
Bill	0	10	0	10	0	10	0	10	1	9	0	10
Bump	0	10	0	10	0	10	1	9	1	9	0	10
Brown	0	10	0	10	0	10	0	10	1	9	0	10
Roger	0	10	0	10	0	10	0	10	1	9	0	10
Blue	0	10	0	10	0	10	0	10	0	10	0	10
White	0	10	0	10	0	10	6	4	5	5	1	9
Green	0	10	0	10	0	10	2	8	9	1	6	4

Table 5–1: Individual binary choice data from *early* phase.

The first study was considered the best starting place because the triadic tests were novel to subjects at this point, and so it can most reasonably be assumed that subjects attempted to apply the the same strategy that was successful for the binary tests. This means that the simple *naive* control strategy (which also makes the most straightforward predictions) is the most appropriate. Later triadic phases present more of a problem for two reasons. First, the case for assuming that subjects continue to use the same set of rules is weaker. Secondly, (b) the *semi-iterative* control strategy needs to be invoked, and this contains an additional variable to cater for (the probability of the control structure iterating). Although the stack model also makes predictions about (relative) reaction times, there is another reason for preferring choice data, in the first instance, which is related to the following problem.

A major difficulty in evaluating the stack model is that there is no way of telling in advance which stack of rules a subject might be using. There are three possible approaches to this problem, all of which will be used in some form or other in this analysis.

1. As we have no prior reason to suppose that any particular stack is more likely to be employed than any other, the predicted profile for a reasonably sized group of subjects would be that formed by taking the average of

the projections from individual rule stacks. Even with a small group of subjects, such as we have with the monkeys, this approach still has some validity. This is because the behaviour of the eight rule stacks overlap to a great extent, and can thus be thought of specifying a characteristic range of behaviour rather than eight discrete categories. This range is reasonably well reflected by the binary enumeration of the stacks from 1 to 8 (*eg* see table 5–3). An 'averaged' stack model should therefore give a good approximation to the averaged data for a group of subjects as long as they employed a reasonably varied sample from the range of stacks.

2. Another approach is to take the eight separate projections from each of the stacks and to compare these with the obtained profiles for individual subjects. This may appear somewhat *post hoc* in that obtaining a degree of fit to the data is inevitably more likely if there are a varied set of projections to choose between. However, this still constitutes a test of the model because the eight profiles cover a restricted range of possible behaviours and, in principle, it is certainly possible for a subject to show a pattern of choices which falls completely outside the space of behaviours allowed by the model. Moreover, when assessing the degree of fit between a projected and obtained profile, the criteria can be adopted that the fit should be (a) better than for the 'averaged' stack model and (b) better than other models.

3. The strongest test of the model is to take a stack which has been assigned to a subject on the basis of one set of data points, and to assess its fit on different data for the same subject. The latter could come from a different block of trials, a different measure (*eg* error *vs* reaction time) or, if the original data points were means, they could be expanded into smaller chunks.

This following sections therefore evaluate the stack model against the triadic choice data, starting with the *early* phase and the *naive* control strategy. The general tactic is to organise both monkey data and the *projected* data (predicted

from the model) into the same form so that they can be easily compared. In doing this, the approach is to chunk the data progressively more finely, starting with a relatively coarse summary and tending towards the level of individual trials. Subjects' performances on the binary tests were generally too good to generate interpretable error data but binary reaction times are dealt with in a later section.

5.1.1 Generating Projections for the Triads

Assuming a *naive* control strategy, triadic choice profiles can be generated for each of the eight stacks as shown in table 5–2 for the example stack [1]. The projection is for a 100 trials, ten for each triad. It was assumed that on triads where the shallowest applicable rule is an avoidance one, the choices are split 50/50 among the remaining items. This would only be true on average and only providing that such uninformed choices are random (as opposed to the subject selecting a favourite colour, for example). The binary sampling model makes an analogous assumption, and the fit of the two models will be compared in this section. The bottom row of the table summarises the projection to five data points (sums), one per choice item, in the same way as was done for the actual group data in table 3–4.

Projections have similarly been made for the other rule stacks and five point summaries of all eight are shown in table 5–3. As the projection is for a total of 100 trials in each case, the figures can conveniently be read as percentages of the total accruing to the corresponding item. Note that the profile for stack 1 is the same as that for perfect performance and that performance degrades in a way correlated with the stack number. This table effectively summarises a space of possible projected profiles, and will be used extensively. The bottom row shows the averaged distribution across all stack forms and is thus a summary of the

---

[1]Exactly the same projection is obtained with the alternative rule 4: $present(B) \Rightarrow avoid(B)$.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

Triad	*A*	*B*	*C*	*D*	*E*
ABC	0	5	5	–	–
BCD	–	0	0	10	–
BDE	–	0	–	0	10
CDE	–	–	0	0	10
BCE	–	0	0	–	10
ABD	0	5	–	5	–
ACD	0	–	5	5	–
ADE	0	–	–	0	10
ABE	0	0	–	–	10
ACE	0	–	0	–	10
Totals	0	10	10	20	60

Projected average distribution of choices to each item with ten presentations of each triad and *naive* control strategy.

Table 5–2: Projected choice distributions for stack 3.

Identity	*RuleForms*	*A*	*B*	*C*	*D*	*E*
1.	*SeSeSe*	0	0	10	30	60
2.	*SeSeAv*	0	5	5	30	60
3.	*SeAvSe*	0	10	10	20	60
4.	*SeAvAv*	0	10	15	15	60
5.	*AvSeSe*	0	15	15	25	45
6.	*AvSeAv*	0	15	20	20	45
7.	*AvAvSe*	0	15	25	25	35
8.	*AvAvAv*	0	15	25	30	30
–	*Overall%*	0	10.6	15.6	24.4	49.4

Table 5–3: Projected percentage of total choices to each item for all stack forms.

'averaged stack model' described in the previous section. Note that this is not the only way of summarising the projections; a 'horizontal' summary (by triad instead of by item) will be made use of further on.

Preliminary comparison with group monkey profile

Two of the monkeys in this sample break a basic assumption of the stack model, which is that each subject possesses a rule stack which is fully functioning for the binary pairs. Two subjects performed erratically on some pairs in the binary testing phase, whilst the remaining five were almost error free. Strictly speaking, therefore, the stack model (in its current form) is only valid for these five subjects. However, previously, the data from all seven subjects has been analysed together, so it is in order to make a comparison of the averaged stack projection with the data grouped as previously published (McGonigle & Chalmers, 1977; McGonigle & Chalmers, 1986).

In table 5–5, the averaged stack profile is compared with the projection from

the binary sampling for fit to the combined profile from seven subjects. The binary sampling model has already been compared with the monkey profile in chapter 3 (last two rows of table 3–4), but an attempt at a more formal analysis is given here, based on the chi-squared goodness-of-fit test. This is a method of comparing a set of observed (independently sampled) frequencies with the set of frequencies expected according to some model. It assumes a binomial distribution of sampling error in the observed frequencies. A theoretical chi-squared distribution is used to assess whether the observed set of frequencies is significantly different from those expected according to the model.

For example, the averaged stack model predicts choices to be distributed, in certain proportions, between the B, C, D and E items. Under the null hypothesis (H_0) that these proportions are correct (*ie* they would arise given an sufficiently large sample size), we can use the chi-squared test to tell us the probability of obtaining the proportions actually observed. The following convention is adopted for interpreting the results. A low probability ($p < 0.05$) means that the H_0 can reasonably be rejected — the model is inaccurate. An intermediate probability ($p > 0.1$) means that there is no reason to reject H_0 — the data lends some support to the theory. High probabilities ($p > 0.5$) mean (in the psychological domain), that the model describes the data well. This categorisation leaves a grey area of probabilities between 0.05 and 0.1, where there there is insufficient data to either accept or reject the model. The significance levels used here relate to the standard ones employed in more commonly used statistical distributions, as shown in table 5–4.

The conclusions of this test can only be accepted with certain reservations for this application. First, the frequencies of choices to the items B, C, D and E are not totally independent. For example, it is impossible for all the choices to go to E with none to the other items. However, bearing in mind the stochastic component of the model, the situation does approximate to the conditions required by the test. Second, the test is based on a null hypothesis which is a little too stringent for psychological models. In rejecting H_0 we should not necessarily reject the model which may in fact provide a useful approximation

Probability (p)	Interpretation
$p > 0.1$	No grounds to reject H_0
$0.05 < p < 0.1$	Insufficient evidence to reject H_0
$p < 0.05$	Reject H_0 (5% level)
$p < 0.01$	Reject H_0 with confidence (1% level)
$p < 0.001$	Reject H_0 (highly significant)

Table 5–4: Standard significance levels and interpretation heuristics (Experimental Psychology).

to the data, *albeit* consistently slightly inaccurate. For this reason we interpret the intermediate probabilities liberally. Finally, we will only be testing for the ability of models to predict proportions of choices to four of the items; they are already correct in predicting zero choices to the A item.

Table 5–5 shows a chi-squared goodness-of-fit test applied to the averaged stack model and the binary sampling model. The measure of total deviation between observed (O) and expected (E) frequencies is given by the formulae:

$$\chi^2 = \sum \frac{(0-E)^2}{E}$$

A zero value χ^2 indicates a perfect fit, whilst higher values can be used to assess the relative fit of different models and can also be interpreted with respect to the theoretical chi-squared distributions. First, it can be seen that the binary sampling model gives a closer fit than the stack model (though both are much much better than a perfectly transitive profile). As both models predict zero choices to A this has to be eliminated from the goodness-of-fit test. As the total number of choices is fixed, the relevant theoretical distribution is chi-squared with 3 degrees of freedom.

It can be seen that the data gives some support to the binary sampling model $(0.2 > p > 0.1)$(whilst the (averaged) stack model can reasonably be rejected $(p < 0.001)$. In particular, the stack model predicts too many choices to the

E item. However, both models provide a much better fit than the perfectly transitive profile (not shown). As already stated, this particular comparison has been made primarily because the monkey data has already been published in this form (though this statistical test has not previously been applied). The binary sampling model also performs somewhat better when the data and projections are summarised by triad, as in the original *Nature* paper (table 3–2).

The subsequent analysis of the *early* phase concentrates on the combined and individual data of the five subjects Bill, Blue, Bump, Brown and Roger, whose performances were all near perfect on the binary tests.

$\ddots$	A	B	C	D	E	$\sum$	$p(Observed)$
Monkeys (O)	1	91	140	180	288	700	–
Stack Mod. (E)	0	74	109	171	346	700	–
$\frac{(O-E)^2}{E}$	–	3.91	8.82	0.47	9.72	$\chi^2 = 22.9$	$p < 0.001$
Bin. Samp. (E)	0	105	154	161	280	700	–
$\frac{(O-E)^2}{E}$	–	1.87	1.27	2.74	0.23	$\chi^2 = 6.1$	$p > 0.1$

The top box (tow rows) shows the *observed* frequencies of choices to each item (*early* phase). The second box down shows the frequencies *expected* according to the averaged stack model, with measures of deviation in the row below. The bottom box (last two rows) similarly shows expected frequencies and deviations for the binary sampling model. The $\sum$ column shows the sums of the entries in the rows to the left. The summed deviances give the 'goodness of fit' statistic, χ^2. Each entry in the column on the far right gives the probability of the of the observed frequencies arising, under the null hypothesis H_0 that there is no significant difference between expected and observed frequencies. The probabilities are obtained from standard chi-squared tables (3 degrees of freedom).

Table 5–5: Relative fits of average stack profile and binary sampling model to grouped data for all seven subjects.

5.2 Study 1

Following the methodology described above, the first step is to evaluate the fit of the averaged stack model to the (redefined) group data with two subjects rejected.

5.2.1 Global Fit

Table 5–6 compares the relative fit of the data to the averaged stack model, binary sampling model and a perfect transitive profile. The format is the same as in table 5–5. It can be seen that, with two subjects removed, the relative fit of the models is strikingly reversed. The data fits the stack model very well ($0.7 > p > 0.5$), whereas the binary sampling model can be safely rejected. The data are also far removed from the perfect response profile.

5.2.2 Microanalysis by Triad

The projections can also be summarised in a different direction, which is to calculate the average of the eight projections for each triad. Table 5–7 compares these with the mean monkey distribution and analogous projections from the binary sampling model. See appendix A for the calculations.

Inspection of this table reveals that the stack model gives a better fit to the data than the binary sampling model, both with respect to the mean distribution of choices between γ, β and α items and at the level of individual triads. In particular, the binary sampling model predicts choices to the γ item on the triads BCD and BDE which are not reflected in the data for this group. The stack model predicts zero choices to the γ item. An important question remains, however. Is the relative success of the stack model entirely due to correctly projecting the *overall* relative proportions of choices to the α and β or is the projected variation *between* the triads significant? In other words, would the fit

$\ddots$	A	B	C	D	E	Σ	$p(Observed)$
Monkeys (O)	1	46	72	131	250	500	–
Stack Mod. (E)	0	53	78	122	247	500	–
$\frac{(O-E)^2}{E}$	–	0.93	0.46	0.66	0.04	$\chi^2 = 2.1$	$p > 0.5$ $(3df)$
Bin. Samp. (E)	0	75	108	117	200	500	–
$\frac{(O-E)^2}{E}$	–	6.45	12.17	1.68	12.50	$\chi^2 = 32.8$	$p \ll 0.001$ $(3df)$
Perfect TI (E)	0	0	50	150	300	500	–
$\frac{(O-E)^2}{E}$	–	–	9.68	2.41	8.33	$\chi^2 = 20.4$	$p \ll 0.001$ $(2df)$

Observed frequencies of choices to each item (*early* phase) compared with those expected according to the (averaged) stack model, the binary sampling model, and a perfectly transitive profile. CF table 5–5.

Table 5–6: Relative fits of average stack profile and binary sampling model to grouped data for five subjects.

be just as good if a 0, 25, 75% distribution were the projection for each individual triad?

If the handful of choices to the γ item are disregarded, then the main determinant of the profiles can be considered to be the proportion of choices to the α item (the proportion choices to the β item can then be calculated by subtraction). The previous question can therefore be answered by finding the statistical correlation between the respective α columns. This is only valid for the stack model, as the binary sampling model predicts too many choices to the γ item. Spearman's correlation statistic indicates a significant positive correlation ($r = 0.634$: see table 5–8) between the ten projected and obtained α values. An interpretation of this result is that the averaged stack model can account for approximately 40% (r^2) of the inter-triad variance for this group of subjects[2]. This is an encouragingly high proportion, considering the low number of sub-

[2]The correlation coefficient r ranges from +1 (maximum correlation) through 0 (no

Triad	*Stack Mod.*			*Monkeys*			*B. Samp.*		
$\gamma\ \beta\ \alpha$	γ	β	α	γ	β	α	γ	β	α
ABC	0	44	56	0	38	62	0	33	67
BCD	0	25	75	4	26	70	17	33	50
BDE	0	12*	87*	0	20	80	17	17	67
CDE	0	6	94	2	18	80	0	33	67
BCE	0	12*	87*	2	12	86	0	33	67
ABD	0	37*	62*	0	34	66	0	50	50
ACD	0	37*	62*	0	22	78	0	50	50
ADE	0	25	75	0	12	88	0	33	67
ABE	0	25	75	0	14	86	0	33	67
ACE	0	25	75	0	20	80	0	33	67
Means	0	25	75	1	22	78	3	35	63

Projection of averaged (*naive*) stack model compared with monkey distributions and binary sampling model projections. All percentages are rounded to nearest whole number. *Rounded down 0.5%.

Table 5–7: Relative fit of averaged stack model at triadic level.

Projected	56	75	87.5	94	87.5	62.5	62.5	75	75	75
Obtained	62	70	80	80	86	66	78	88	86	80

Projected and obtained choices to α item (from table 5–7). Spearman's product-moment correlation coefficient: $r = 0.634$ (t test $p < 0.05$).

Table 5–8: Correlation of averaged stack model with group data

jects and the fact that, according to the model, we expect differences between individuals to contribute to the overall variance. Also, some stochastic variance would be expected due to the random component of the control strategy.

The overall conclusion from the two previous analyses is that the (averaged) stack model gives a very promising degree of fit to the data from the five best subjects. It seems that the binary sampling model would need to be significantly modified to account for the data from these subjects. Nevertheless it will be retained as a useful yardstick in subsequent analysis. Also, the possibility that some individuals might be using a binary sampling strategy should not be ruled out at this stage. This brings us on to the question of why the averaged stack model provides such a good fit. It could be that the assumptions behind it are correct and that the five individuals are each using a rule stack and that the sample of stacks is reasonably spread across the eight possibilities. This possibility is explored below. On the other hand, it could be that the averaged stack model is successful for different (unanticipated) reasons and that the averaged profile would turn out to be accurate at the individual as well as the group level.

5.2.3 Microanalysis by Individual

It ought to be possible to account for more of the group variance in the previous analysis if it could be discovered which five rule stacks were being employed, rather than averaging the projections from all eight stacks. The next step, therefore, is to attempt to assign rule stacks to individual subjects. If the model is correct it should be possible, in principle, to obtain a much closer fit at the individual level. However, the problem is that there is also much less data at the individual level. At the group level there are 50 trials for each triad and so random components of the decision process would tend to average out. At the

---

systematic relationship between variables) to -1 (maximum negative correlation). Correlation and its relationship with variance are described in standard statistics references.

individual level there are only ten trials per triad so random effects can have a severe effect of the distribution of choices between the critical α, β items.

For this reason, it was decided to take the same approach as for the group data, and summarise each subject's choices to a five point profile. Each could then be compared with the eight projected summaries shown in table 5–3 to find the best match. In each case the best match ought to fit better than the averaged stack model. The results are shown in tables 5–9 to 5–13 and are commented individually.

Overall, the analyses provide strong supportive evidence for the stack model. It may be concluded that the stack model is an essentially correct description of five subjects' decision procedures in the *early* testing phase, although some of the simplifying assumptions may be incorrect in detail. In particular, the assumption about individual's rule stacks being invariant may be incorrect for at least one subject, namely Blue.

$\ddots$	A	B	C	D	E	Σ	$p(Observed)$
Bill	0	8	20	17	55	100	–
Stack 4.	0	10	15	15	60	100	–
$\frac{(O-E)^2}{E}$	–	0.40	1.67	0.60	0.42	$\chi^2 = 3.1$	$p > 0.3$
A.S.P	0	11	16	24	49	100	–
$\frac{(O-E)^2}{E}$	–	0.82	1.00	2.04	0.51	$\chi^2 = 4.4$	$p > 0.2$
B.S.M.	0	15	22	23	40	100	–
$\frac{(O-E)^2}{E}$	–	3.27	0.18	1.57	5.63	$\chi^2 = 10.7$	$p < 0.02$

This and the following four tables have the same structure as table 5–5. The abbreviations are for the *A*verage *S*tack *P*rofile and the *B*inary *S*ampling *M*odel.

Comment: Although this subjects data fits the averaged stack model well, stack 4. provides an even better fit. The subject does not appear to be employing a binary sampling strategy. This is supportive evidence for the stack model account.

Table 5–9: Bill

$\ddots$	A	B	C	D	E	Σ	$p(Observed)$
Blue	0	6	14	25	55	100	–
Stack 3.	0	10	10	20	60	100	–
$\frac{(O-E)^2}{E}$	–	1.60	1.60	1.25	0.42	$\chi^2 = 4.9$	$p > 0.1$
A.S.P	0	11	16	24	49	100	–
$\frac{(O-E)^2}{E}$	–	2.27	0.25	0.04	0.73	$\chi^2 = 3.3$	$p > 0.3$
B.S.M.	0	15	22	23	40	100	–
$\frac{(O-E)^2}{E}$	–	5.40	2.91	0.17	5.63	$\chi^2 = 14.1$	$p < 0.01$

Comment: None of the eight stacks fit better than the averaged stack model for this subject. Nevertheless, stack 3 cannot be rejected as a model of the data ($p > 0.1$). The binary sampling model can be rejected. That the averaged stack profile fits so well, suggests that the subject is using a stack of rules which is changing over the course of the test or that the subject is using something other than a *naive* control strategy. Either solution breaks some assumptions of the stack model.

Table 5–10: Blue

$\ddots$	A	B	C	D	E	Σ	$p(Observed)$
Bump	1	4	8	34	53	100	–
Stack 2.	0	5	5	30	60	100	–
$\frac{(O-E)^2}{E}$	–	0.20	1.80	0.53	1.07	$\chi^2 = 3.6$	$p > 0.3$
A.S.P	0	11	16	24	49	100	–
$\frac{(O-E)^2}{E}$	–	4.45	4.00	4.17	0.33	$\chi^2 = 12.9$	$p < 0.01$
B.S.M.	0	15	22	23	40	100	–
$\frac{(O-E)^2}{E}$	–	8.07	8.91	5.26	0.23	$\chi^2 = 22.5$	$p < 0.001$

Comment: This subject performs exceedingly well on the triads and the perfect transitive profile (stack 1) fits the data well. However, the stack 2 profile provides an even better characterisation. Both the other models, and the remaining stacks, can be rejected. This is strong supportive evidence for the stack model account.

Table 5–11: Bump

⋱	A	B	C	D	E	Σ	$p(Observed)$
Brown	0	11	24	29	36	100	–
Stack 7.	0	15	25	25	35	100	–
$\frac{(O-E)^2}{E}$	–	1.07	0.04	0.64	0.03	$\chi^2 = 1.8$	$p > 0.5$
A.S.P	0	11	16	24	49	100	–
$\frac{(O-E)^2}{E}$	–	0	4.00	1.04	3.45	$\chi^2 = 8.5$	$p < 0.05$
B.S.M.	0	15	22	23	40	100	–
$\frac{(O-E)^2}{E}$	–	1.07	0.18	1.57	0.40	$\chi^2 = 3.2$	$p > 0.3$

Comment: Unlike Bump, this subject performs poorly on the triads and stack 7 (the second worst) gives a very good fit to the data ($p > 0.5$). The binary sampling model also gives a (less) good fit in this case but the averaged stack profile can be rejected. This analysis provides important supportive evidence for the stack model account because it shows that the improved accuracy of a single stack over other models is not limited to subjects with a strong transitive bias on the triads.

Table 5–12: Brown

⋱	A	B	C	D	E	Σ	$p(Observed)$
Roger	0	17	6	26	51	100	–
Stack 5.	0	15	15	25	45	100	–
$\frac{(O-E)^2}{E}$	–	0.27	5.40	0.04	0.03	$\chi^2 = 6.5$	$0.05 < p < 0.1$
A.S.P	0	11	16	24	49	100	–
$\frac{(O-E)^2}{E}$	–	3.27	6.25	0.04	0.08	$\chi^2 = 9.7$	$p < 0.05$
B.S.M.	0	15	22	23	40	100	–
$\frac{(O-E)^2}{E}$	–	0.27	11.6	0.39	3.01	$\chi^2 = 15.3$	$p < 0.01$

Comment: In one respect, the data from this subject runs counter to all the models (and to the data from the other subjects) in that there appear to be significantly more choices to B than to C. Although none of the models actually rules out such an outcome, the averaged stack profile and the binary sampling model are sufficiently deviant to be rejected. The best fitting stack is No. 5 which is not conclusively supported or rejected by the data. Overall, the data form this subject supports (*albeit* weakly) the stack account. It is a reasonable hypothesis that there was a chance perturbation in the data which a greater number of trials would have ironed out. However, it may be the case that one or more of our assumptions are wrong, as was suggested for Blue.

Table 5–13: Roger

5.2.4 Microanalysis by Individual and by Triad

Following the same methodology as for the group data, the next step is to break down the projections for individuals into individual triads. Tables 5–14 to 5–18 show the stacks selected for each subject (above) with the projected and obtained distributions for each triad. Only choices to the α and β items are shown. The number of γ choices can be obtained by subtracting the α and β choices away from 10 (the total number of trials for each triad). In the case of the stack projections, this is always zero, but a scattering of choices do go to γ items for the subjects Bump, Brown and Roger. This is just over 1% of all choices.

Stack 4.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(B) \Rightarrow avoid(B)$
4) $present(D) \Rightarrow select(D)$

Triad	*Stk 4.*		*Fit*	*Bill*	
$\gamma\ \beta\ \alpha$	β	α	$\times\surd$	β	α
ABC	5	5	$\times$	1	9
BCD	5	5	$\surd$	6	4
BDE	0	10	$\surd$	0	10
CDE	0	10	$\surd$	1	9
BCE	0	10	$\surd$	1	9
ABD	5	5	$\surd$	5	5
ACD	5	5	$\surd$	3	7
ADE	0	10	$\surd$	0	10
ABE	0	10	$\surd$	2	8
ACE	0	10	$\surd$	1	9
Totals	20	80	9	20	80

Comment: Good overall fit except for the triad ABC ($p \approx 0.01$).

Table 5–14: Triads: Bill

Stack 3.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

Triad	*Stk* 3.		*Fit*	*Blue*	
$\gamma\ \beta\ \alpha$	β	α	$\times\surd$	β	α
ABC	5	5	$\times$	2	8
BCD	0	10	?	3	7
BDE	0	10	$\surd$	1	9
CDE	0	10	$\surd$	0	10
BCE	0	10	$\surd$	0	10
ABD	5	5	$\surd$	4	6
ACD	5	5	$\times$	1	9
ADE	0	10	$\surd$	2	8
ABE	0	10	$\surd$	0	10
ACE	0	10	$\surd$	2	8
Totals	15	85	7	15	85

Comment: Poor overall fit ($p \approx 0.17$) especially for the triad BCD, which is worse than expected, and the triad ACD which is better than expected.

Table 5–15: Triads: Blue

Stack 2.

1) $present(E) \Rightarrow select(E)$
2) $present(D) \Rightarrow select(D)$
3) $present(A) \Rightarrow avoid(A)$
4) $present(C) \Rightarrow select(C)$

Triad	*Stk 2.*		*Fit*	*Bump*	
$\gamma\ \beta\ \alpha$	β	α	$\times\surd$	β	α
ABC	5	5	$\surd$	3	7
BCD	0	10	$\surd$	0	10
BDE	0	10	$\surd$	2	8
CDE	0	10	$\surd$	2	8
BCE	0	10	$\surd$	0	10
ABD	0	10	$\surd$	0	10
ACD	0	10	$\surd$	0	10
ADE	0	10	$\surd$	0	9
ABE	0	10	$\surd$	0	9
ACE	0	10	$\surd$	1	9
Totals	5	95	10	8	90

Comment: Highly significant overall fit ($p \approx 0.001$) Performance is slightly worse than expected on triads BDE and CDE. The performance on ABC is better than expected but still within chance levels.

Table 5–16: Triads: Bump

Stack 7.

1) $present(A) \Rightarrow avoid(A)$
2) $present(B) \Rightarrow avoid(B)$
3) $present(E) \Rightarrow select(E)$
4) $present(D) \Rightarrow select(D)$

Stack 8.

1) $present(A) \Rightarrow avoid(A)$
2) $present(B) \Rightarrow avoid(B)$
3) $present(C) \Rightarrow avoid(C)$
4) $present(E) \Rightarrow select(E)$

Triad	*Stk 7.*		*Fit*	*Brown*	
$\gamma\ \beta\ \alpha$	β	α	$\times\surd$	β	α
ABC	5	5	$\surd$	4	6
BCD	5	5	$\surd$	3	6
BDE	5	5	$\surd$	5	5
CDE	0	10	$\times$	5	4
BCE	5	5	$\surd$	5	5
ABD	5	5	$\surd$	4	6
ACD	5	5	$\surd$	5	5
ADE	5	5	$\times$	2	8
ABE	5	5	$\times$	2	8
ACE	5	5	$\surd$	4	6
Totals	45	55	7	39	59

Comment: Stack 7 narrowly misses achieving a 5% significance level. The main problem is with the triad CDE, which is worse than expected. Stack 8, which projects a 0, 5, 5 distribution to all triads would have provided a significant fit. Another mismatch is that the triads ADE and ABE are more biased than expected. Looking at the overall proportions obtained (39/59%), it seems that the hypothesis that the distribution is 50/50% on *random* triads is probably incorrect. This might be explained if the subject were using something better than a *naive* control strategy.

Table 5–17: Triads: Brown

Stack 5.

1) $present(A) \Rightarrow avoid(A)$
2) $present(E) \Rightarrow select(E)$
3) $present(B) \Rightarrow avoid(B)$
4) $present(D) \Rightarrow select(D)$

Stack 3.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

Triad	*Stk* 5.		*Fit*	*Roger*		*Stk* 3		*Fit*
$\gamma\ \beta\ \alpha$	β	α	$\times\surd$	β	α	β	α	$\times\surd$
ABC	5	5	×	9	1	5	5	×
BCD	5	5	×	1	8	0	10	√
BDE	0	10	√	2	8	0	10	√
CDE	0	10	√	1	9	0	10	√
BCE	0	10	√	0	9	0	10	√
ABD	5	5	√	4	6	5	5	√
ACD	5	5	?	2	8	5	5	?
ADE	5	5	×	1	9	0	10	√
ABE	5	5	?	2	8	0	10	√
ACE	5	5	?	2	8	0	10	√
Totals	30	70	4	24	74	15	85	8

Comment: The stack 5 projection does not give a good fit to the individual triads. In particular, the peculiar distribution of choices to the triad ABC is not predicted by any version of the model. Extra choices also went to B on the triads BCD and BCE suggesting that the subject had some kind of bias in favour of this item. This suggests that the selection of stack 5 was an anomaly. If the apparent bias towards B is compensated for, then the best fitting stack is number 3, as shown on the right, which would have achieved a 1% significance level.

Table 5–18: Triads: Roger

Even a cursory comparison of expected and obtained numbers of choices shows that the stack model is not literally correct. There are only two kinds of distributions projected, those where 100% of choices are expected to go to the α item (**alpha triads**), and those where a random (50/50) distribution is expected between β and γ items (**random triads**). Consider first the alpha triads. Overall, 10.8% of the available choices on such trials are diverted to β and γ items. This seems too high a percentage to be simply attributed to noise. Furthermore, it can be broken down into 9.6% diverted to the β item and only 1.2% to γ. If these α choices were being lost because of some completely random noise element we would expect them to be diverted equally between the remaining alternatives. Instead (in terms of the model) it is as though subjects occasionally fail to identify the shallowest applicable rule in the stack and go on to select the second best choice. Another possibility is that the rule order itself is not rigid but is subject to some variation (see section 4.6.1).

These possible modifications to the stack model suggested above would also help explain the unexpectedly high degree of fit obtained with the averaged stack model; if the rule order is not rigid then one stack may temporarily appear like another. However, before constructing a more complex model it is still worth assessing how good the simple stack model is as an approximation to the decision mechanism, even though it is wrong in detail. The basic problem is to see if the data reflects the categorisation of triads as either *alpha* or *random*. In order to do this we need an independent criterion for classifying the choice data.

We will, therefore, consider the data to be an adequate fit to the projection on alpha triads if 80% or more of the choices go to the α item (8/10). This threshold is low enough to catch the alpha triads and (just) high enough to exclude most random triads. According to probability theory, there is only a 5.5% ($p < 0.055$) chance that eight or more of the ten choices would go to the α item (given the null hypothesis that the probability is 0.5 on each trial). Unfortunately, this gives us a 5% chance of misclassifying a random triad as an alpha triad. The chance of misclassifying an alpha triad as a random one is unknown, but is hopefully no higher. Triads will be classified as random if the split of choices between α

and β items is no more skewed than 30/70% in either direction. 89% of random triads should have a distribution in this range ($3 < \alpha < 7$), according to the null hypothesis. Each of the tables (5–14 to 5–18) includes a column of ticks and crosses. A tick indicates a correspondence between the projection and the data according to the above classification. A question mark indicates a narrow miss (8/10 for a random triad and 7/10 for an alpha triad. The number of ticks is summed at the bottom, thus giving an overall measure of fit.

Taking into account the above assumptions and the inherent problem of misclassification, a score of 8/10 or more should be expected, for an individual, if the stack model is a good approximation and the particular stack selected is appropriate. To see why this is the case, suppose we take a null hypothesis that both the monkey data and the projection consist of a random mix of *alpha* and *random* triad types (ignoring the possibility of triads biased heavily towards the β item). This is a good null hypothesis as it is compatible with the overall observed distribution. It leads to a 0, 25, 75% distribution, as does the averaged stack model. According to the hypothesis, the chance of getting the projection correct for any particular triad is 50%. Thus a score of 5/10 would mean that the fit was no better than chance level. The probability of getting 8/10 (or better) correct is about 5%, the probability of 9+/10 is about 1% and the probability of getting 10/10 correct is about 0.1% (probabilities from binomial distribution, N=10). A score of 7/10 may mean that the stack has some predictive value but there is not enough data to reject the null hypothesis for that individual.

Conclusions

Overall, the combined score for the five subjects is 37/50 for the original five stacks (selected in section 5.2.3). With this combined score the null hypothesis can be confidently rejected at the 0.1% significance level ($p < 0.001$ using a normal approximation to the binomial curve, N=50). This means that the stack model has some value as an approximation to the data for individuals — above and beyond predicting the correct overall choice proportions to the γ, β and α items (0, 25, 75%).

The fit is not significant for every individual, however. Two subjects, Bill and Bump (with scores of 9 and 10, respectively), unambiguously pass the 8/10 criterion. Brown narrowly misses (see commentary in table 5–17), and would have passed with stack 8, which differs from the selected stack, 7, on one triad. It seems that the selection of stack 5 for Roger was a mistake due to an unexpected bias towards the item *B* which distorted the summarised choice profile (table 5–13). Stack 5 was not conclusively accepted previously (table 5–13) and stack 3 gives a much better fit at the local level. There is no stack which gives a better fit for Blue on the other hand, and it may be that this subject is doing something different from the others. This fits in with the previous finding (table 5–10) that no individual stack could be found which fitted better than the averaged stack model.

The overall score is 42/50 with the *post hoc* replacement stacks for Brown and Roger. The significance of this is that five stacks can be matched up to individuals such that 84% of triads can be correctly classified as belonging to one of two categories:

alpha triads in which 80% or more of the choices go to the α item.

random triads in which choices are split between the *alpha* and β items in a ratio not exceeding 70/30%.

This 84% is almost as high a percentage as could maximally be expected, given that about 5% of *random* would be misclassified as *alpha* and that some misclassification in the other direction is also possible.

Given the partial success of this classification system, under the final assignment of rule stacks, it makes sense to reassess the assumptions behind the *naive* control strategy. If we combine the data from all the triads projected as *random* and *alpha* separately, we can assess whether they are significantly different in the actual data. This is shown in table 5–19. It can be seen that the relative proportions of choices to β and α items are 39/61% and differ significantly from the expected 50/50% distribution. Thus it may be the case that the *semi-iterative*

control strategy (see previous chapter) is more appropriate. For example, If there was approximately a 20% chance of the control strategy iterating then this would give a 40/60% distribution on the *random* triads.

$\ddots$	β	α	No. Triads
Random	3.9	6.1	21
Alpha	1.1	8.9	29

Significantly more than half the available choices go to the α item in both types of triad (binomial test $p < 0.001$). However, significantly more choices go to the α item in *alpha* triads than in *random* triads (t test for independent samples).

Table 5–19: Mean distributions for *random* and *alpha* type triads.

5.2.5 Summary

1. Theoretical choice distributions were generated for the eight significantly different rule stacks assuming a *naive* control strategy. These were then summarised in various ways for ease of comparison with the *early* triadic choice data. In particular, the average profile of all eight stacks was generated for comparison with the group data.

2. The fit of the averaged stack model profile was compared with the averaged data for five subjects and was found to fit very well.

3. Rule stacks were selected for individual subjects by finding the best match from eight projected profiles to five point summaries of each subject's triadic choice data. The degree of fit obtained varied but was better than the fit to the binary sampling model and the perfect transitive profile in all cases. In all but one case, a better fit could be found with an individual stack than for the averaged stack model. This supports the contention that the success of the averaged stack model is due to the five individuals using different rule stacks.

4. The projected profiles from the five selected stacks were then tested for fit against the individual data on a triad by triad basis. It was found that the stack model failed to predict a significant diversion of choices away from the α items to the β items. It was suggested that the assumption of a rigid rule order for each individual may be incorrect, in that there may be some fluctuation.

5. The degree to which the individual stack projections successfully approximated the subjects data was assessed. Overall, the stack model did significantly better than chance, although there appeared to be one or two errors in the previous selection of stacks for individuals. In the final analysis, it was found that four out of the five subjects could be assigned rule stacks which gave a convincing degree of fit to their data. One subject (Blue) appeared to differ from the others, and this may have been due to (a) the rule order changing during the course of the test or (b) the subject using a different control strategy.

6. The data from all the triads which were projected as random were combined and it was found that approximately 60% (significantly more than half) of the choices went to the α item. This suggests that the control strategy was iterating on approximately 20% of trials.

The overall conclusion from the analysis presented thus far is that the the stack model with a *naive* control strategy is useful as an approximation to monkeys' decision procedures during the *early* triadic tests. The assignment of different rule stacks to each individual has at least partial validity, although rule order may not be as rigid as anticipated for some subjects.

5.3 Study 2

Five of the subjects in the first study were retrained on the same series as they had previously learned. These were Blue, Brown, Roger, White and Green. White and Green were previously rejected from the analysis but all subjects were fully transitive on the pairs after retraining. Blue, Brown and Roger are analysed with respect to the stacks assigned to them in the previous analysis, on the assumption that these would be retained. White and Green were assigned stacks on the basis of a preliminary analysis of their binary reaction times, explained below.

As all subjects had prior experience of triadic tests, it could not assumed that subjects would be employing a *naive* control strategy. Also, there were two phases of triadic tests in this study, *middle* and *late*, with better performance in the more intensive *late* phase (figure 3–3). It was decided to see if the choice patterns and improvement could be accounted for with the *semi-iterative* control strategy, with different levels of iteration allowed for different subjects and different phases. For each individual, and for each testing phase, the assumed level of iteration will be calculated on the basis of the average bias to the α item on the projected *random* triads.

For the subjects Blue, Brown and Roger, the data from the *early* phase (study 1) is also incorporated into the analysis, to see if the *semi-iterative* control strategy could better account for the choices in this phase too. However, any improvement between the *early* and *middle* phases may be due to the retraining.

5.3.1 Construction of Tables

Tables 5–20, 5–21 and 5–23 have the same basic structure as the previous ones for individual triads except that they show three separate sets of data, *early*, *middle* and *late*. As before, only the choices to α and β items are shown as these add up to nearly the total number of choices. The *total* row shows the percentages

of choices to an item out of the total number of choices (in the corresponding phase). The three 'model' columns show projections from the same stack with differing levels of iteration. The level of iteration was determined as follows. For each phase the average distribution on *random* triads was determined. Here, *random* triads are those predicted to contain a random component in the decision process because the shallowest applicable rule in the stack is an avoidance one. A level of iteration is then chosen to produce the same average distribution in the projected *random* triads.

For example, there are three *random* type triads projected for Blue from stack 3. The average distribution on these in the early phase is 0, 23, 77% to γ, β and α items respectively. According to the *semi-iterative* control strategy, this would mean that 46% $(23 + 23)$ of choices on these triads were split (on average) equally between β and α items and that the remaining 54% $(77 - 23)$ are made correctly to the α item by the control iterating and applying a second rule. The average expected distribution over ten trials is 2.3 to the β item and 7.7 to α.

The bottom row shows (where applicable) the choice distributions on the remaining (*alpha* type) triads. If our previous analysis is correct, this gives an indication of how 'rigidly' the rules are ordered. The more choices that 'leak' to the β item on these triads, the less tenable the assumption of a fixed stack order is and, also, the more difficult it is to distinguish *alpha* and *random* type triads empirically. In the previous example, the proportions are 11% to β and 89% to α in the early phase, which is difficult to distinguish from the 23/77% proportions on *random* triads. In contrast, the same ratios for Roger are 13/84% and 50/50%, respectively, which are more likely to be distinguishable. The same problem does not exist for Brown, as no *alpha* type triads are predicted.

5.3.2 Evaluation

Blue

The data for this subject are included for completeness and do not initially appear very interesting at the level of individual triads (see commentary in ta-

Stack 3.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

Triad	*Early*		54%		*Middle*		67%		*Late*		92%	
$\gamma\,\beta\,\alpha$	β	α	β	α	β	α	β	α	β	α	β	α
ABC	2	8	2.3	7.7	2	4	1	5	3	21	1	23
BCD	3	7	0	10	1	5	0	6	0	24	0	24
BDE	1	9	0	10	0	6	0	6	0	23	0	24
CDE	0	10	0	10	0	6	0	6	0	24	0	24
BCE	0	10	0	10	1	5	0	6	0	24	0	24
ABD	4	6	2.3	7.7	1	5	1	5	0	24	1	23
ACD	1	9	2.3	7.7	0	6	1	5	0	24	1	23
ADE	2	8	0	10	1	5	0	6	0	24	0	24
ABE	0	10	0	10	1	5	0	6	0	24	0	24
ACE	2	8	0	10	1	5	0	6	0	24	0	24
Total%	15	85	7	93	13	87	5	95	1	98	1	99
Random%	23	77	23	77	17	83	17	83	4	96	4	96
Alpha%	11	89	0	100	12	88	0	100	1	99	0	100

Comment: There is an improvement in performance for both *random* and *alpha* type triads across the three phases. In each phase there is little difference in the obtained proportions between the two types of triad so it is very difficult to assess the fit of stack 3, especially for the *middle late* phases. These data are therefore not incompatible with the stack model but does not conclusively support it either.

Table 5–20: Three phases: Blue

ble 5–20). From the averaged distributions (bottom three rows) it can be seen that increasing iteration in the control strategy can only account for part of the total improvement. *Random* type triads show a greater improvement in performance than *alpha* ones but this may be simply because there is little room for improvement on alpha triads (a ceiling effect).

Brown

Table 5–21 shows how the gradual improvement across the three phases can be approximately modelled by simply changing the level of iteration. However, a curious feature of this subject's performance is the apparent *deterioration* in the level of choice to the α item on triad ADE from the *early* to later phases. Furthermore, there is a distinct lack of improvement in the triads BDE and CDE. The three 'odd' triads are distinguished by containing both D and E. The most economical explanation is that, for some reason, this subject has dropped the fourth rule ($present(E) \Rightarrow select(E)$) from the stack. Table 5–22 shows how an even better fit to the data can be obtained with a three rule stack with 67% iteration. The fit is only shown explicitly for the *late* phase where it is especially striking. This is probably because there are a larger number of trials per triad (24) and so choice distributions can be estimated more accurately.

This raises the question of why this subject might drop a rule which was (presumably) essential in its binary performance. The fact is that, for any stack, if there is no iteration in the employment of the stack (*ie* a *naive* strategy is used) then the fourth rule is never encountered. A possible explanation is therefore that Brown, after doing a number of trials on the triadic tests, dropped the fourth rule before discovering the need for iteration. It seems feasible that without explicit feedback, the fourth rule would not be regained.

In conclusion, although a certain amount of *post hoc* 'tailoring' has been involved in achieving this degree of fit, the resulting three rule model was thought to be worth including at this point because of its simplicity and plausibility.

Stack 8.

1) $present(A) \Rightarrow avoid(A)$
2) $present(B) \Rightarrow avoid(B)$
3) $present(C) \Rightarrow avoid(C)$
4) $present(E) \Rightarrow select(E)$

Triad	*Early*		20%		*Middle*		33%		*Late*		46%	
$\gamma\ \beta\ \alpha$	β	α	β	α	β	α	β	α	β	α	β	α
ABC	4	6	4	6	1	5	2	4	1	23	7	17
BCD	3	6	4	6	2	3	2	4	4	18	7	17
BDE	5	5	4	6	3	2	2	4	12	12	7	17
CDE	5	4	4	6	2	4	2	4	10	12	7	17
BCE	5	5	4	6	2	4	2	4	7	17	7	17
ABD	4	6	4	6	4	2	2	4	6	18	7	17
ACD	5	5	4	6	2	4	2	4	6	18	7	17
ADE	2	8	4	6	3	3	2	4	10	14	7	17
ABE	2	8	4	6	1	5	2	4	1	23	7	17
ACE	4	6	4	6	1	5	2	4	6	18	7	17
Total%	39	59	40	60	35	62	33	67	26	72	26	72

Comment: No *alpha* type triads are projected for this subject so all the improvement across the three phases is accounted for by increasing iteration. The overall fit is good but note the apparently random choice on triads BDE, CDE and ADE in the *late* phase and, possibly, in earlier phases.

Table 5–21: Three phases: Brown

1) $present(A) \Rightarrow avoid(A)$
2) $present(B) \Rightarrow avoid(B)$
3) $present(C) \Rightarrow avoid(C)$

Triad	*Late*		67%	
$\gamma\ \beta\ \alpha$	β	α	β	α
ABC	1	23	4	20
BCD	4	18	4	20
BDE	12	12	12	12
CDE	10	12	12	12
BCE	7	17	4	20
ABD	6	18	4	20
ACD	6	18	4	20
ADE	10	14	12	12
ABE	1	23	4	20
ACE	6	18	4	20
Total%	26	72	27	71

Comment: The correspondence between the projected and obtained profiles is very close (cf. table 5–21). The 4/20 distribution, projected for most triads, is made up of 8 choices split randomly between β and α items with the remaining 16 going exclusively to α. All the obtained distributions are well within expected chance deviations away from this projection.

Table 5–22: Modelling Brown's *late* phase with a three rule stack

5.3.3 Roger

Table 5–23 shows that the stack model accounts for Roger's data extremely well, with a good fit being obtained on all phases. The previous conclusion about the *early* phase, that the choice distribution on the triad ABC must be some kind of artifact, appears to be vindicated by the consistency in the later phases. The only way in which the data differs significantly from the projections is in that some of the observed improvement is due to better performance on *alpha* type triads. However, over half the overall improvement can be accounted for by the combined improvement of the three *random* triads (this proportion is based on the increases in the number of choices to the α item from the *early* to the *middle* phase). As with Blue, this is not too surprising, as there is more room for improvement on *random* triads.

Stack 3.

1) $present(E) \Rightarrow select(E)$

2) $present(A) \Rightarrow avoid(A)$

3) $present(D) \Rightarrow select(D)$

4) $present(C) \Rightarrow select(C)$

Triad	*Early*		0%		*Middle*		44%		*Late*		44%	
$\gamma\ \beta\ \alpha$	β	α	β	α	β	α	β	α	β	α	β	α
ABC	9	1	5	5	2	4	1.7	4.3	5	19	7	17
BCD	1	8	0	10	2	3	0	6	3	16	0	24
BDE	2	8	0	10	0	6	0	6	0	20	0	24
CDE	1	9	0	10	0	6	0	6	3	21	0	24
BCE	0	9	0	10	0	6	0	6	1	22	0	24
ABD	4	6	5	5	1	5	1.7	4.3	9	15	7	17
ACD	2	8	5	5	2	4	1.7	4.3	6	18	7	17
ADE	1	9	0	10	0	6	0	6	0	24	0	24
ABE	2	8	0	10	0	6	0	6	0	24	0	24
ACE	2	8	0	10	0	6	0	6	0	24	0	24
Total%	24	74	7	93	12	87	9	91	11	85	9	91
Random%	50	50	50	50	28	72	28	72	28	72	28	72
Alpha%	13	84	0	100	5	93	0	100	4	90	0	100

Comment:the inter-triadic variation on all three phases is modelled convincingly by stack 3, with iteration increasing from 0% (a *naive* control strategy) to 44% in the *middle* phase. It seems that performance does not improve at all after the *middle* phase.

Table 5–23: Three phases: Roger

White and Green

These subjects were excluded from the analysis of the *early* phase because of their erratic performance, particularly Green's. Whilst their performance improved after retraining, single stacks could not be unambiguously assigned by the previous method of using a chi-squared goodness-of-fit test. During the *late* phase, on the other hand, their performances were too good for our method of selecting stacks to be readily applicable. It was therefore decided to defer analysis of these subjects choice data until some other method of assigning a stack had been developed. This turned out to be an preliminary analysis of the binary reaction time data, a more rigorous version of which is described in the next section. The mean RTs for each pair were plotted in order to see (by visual inspection) how the RTs for different pairs appeared to group together. It appeared that, for both subjects, a plot according to stack 4 produced the best chunking of RTs and the most linear relationship between depth of rule and RT. A linear regression (on the mean RTs for each pair) confirmed the existence of a strong linear component in the relationship between RT and depth.

Tables 5–24 and 5–25 therefore show choice data from the *middle* and *late* phases for these subjects, analysed in the same way as for the other subjects. Stack 4 fits the data for the *late* phases of both subjects well (and better than other stacks). Green appears to be behaving erratically during the *middle* phase and this may be a hangover from the first study. This was the only subject which originally failed to be transitive on the crucial BD pair and showed a reverse bias on the pair BE.

Overall conclusions

Overall, these results support the stack model. The idea that subjects start with a *naive* control strategy and then begin to increase iteration (*semi-iterative* control) with experience, appears to account for many features of the data. The analysis of the transitions across three testing phases gives further support to the categorisation of triads into *random* and *alpha* types. The main way in which

Stack 4.

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(B) \Rightarrow avoid(B)$
4) $present(D) \Rightarrow select(D)$

Triad	*Middle*		50%		*Late*		50%	
$\gamma\ \beta\ \alpha$	β	α	β	α	β	α	β	α
ABC	1	5	1.5	4.5	2	22	6	18
BCD	2	4	1.5	4.5	8	16	6	18
BDE	1	5	0	6	1	23	0	24
CDE	0	6	0	6	3	21	0	24
BCE	0	6	0	6	1	23	0	24
ABD	2	4	1.5	4.5	9	15	6	18
ACD	1	5	1.5	4.5	4	20	6	18
ADE	0	6	0	6	0	24	0	24
ABE	0	6	0	6	1	23	0	24
ACE	2	4	0	6	0	24	0	24
Total%	15	85	10	90	12	88	10	90
Random%	25	75	25	75	24	76	25	75
Alpha%	8	92	0	100	4	96	0	100

Comment: The data are consistent with the use of stack 4 with 50% iteration for both *middle* and *late* phases. The subject does not appear to improve performance significantly between the phases on either type of triad.

Table 5–24: Two phases: White

Stack 4.

1) $present(E) \Rightarrow select(E)$

2) $present(A) \Rightarrow avoid(A)$

3) $present(B) \Rightarrow avoid(B)$

4) $present(D) \Rightarrow select(D)$

Triad	*Middle*		37%		*Late*		50%	
$\gamma\ \beta\ \alpha$	β	α	β	α	β	α	β	α
ABC	3	2	1.9	4.1	5	15	6	18
BCD	0	5	1.9	4.1	5	17	6	18
BDE	3	3	0	6	1	21	0	24
CDE	2	4	0	6	1	23	0	24
BCE	1	5	0	6	0	24	0	24
ABD	3	3	1.9	4.1	9	15	6	18
ACD	1	4	1.9	4.1	2	22	6	18
ADE	1	5	0	6	0	24	0	24
ABE	3	3	0	6	1	23	0	24
ACE	0	6	0	6	0	24	0	24
Total%	32	67	13	87	10	87	10	90
Random%	29	63	32	68	22	72	25	75
Alpha%	28	72	0	100	2	97	0	100

Comment: There is not much evidence here that this subject is employing stack 4 in the *middle* phase — the data appear very erratic and there is no appreciable difference between *random* and *alpha* types of triad. However, the performance during the *late* phase is much more regular and, with the possible exception of *ACD*, the data fit the stack model well.

Table 5–25: Two phases: Green

the model breaks down (and which was already recognised to be a problem from the analysis of the *early* phase) is that part of the reason for poor performance on *alpha* triads is an apparent loss of choices from α to β items. For Blue and Roger, the observed improvement in performance is partially due to a reduction of this loss. This aspect of the improvement could potentially be accounted for by a model in which subject's stacks became more rigidly ordered with usage, starting out with some kind of 'noisy' or 'dynamic' ordering as suggested in section 4.6.1.

5.4 Study 2: Analysis of Binary RTs

The subsequent analysis is based on binary tests given just prior to the *middle* and *late* triadic tests, and includes the five subjects analysed above. The data consists of the last 100 observations for each subject (10 per pair). There is also an extra subject, Bump, for which RTs were taken a year after the *early* phase as part of a pilot study. The data from this subject are something of a bonus, and they are not subjected to the same in-depth analysis as for the other five subjects.

The procedure adopted was to plot the RTs against 'depth of rule', in order to test the *grouping* and *linearity* predictions described in the previous chapter (section 4.5.1). Both predictions rely on the assignment of a particular rule stack to a subject.

- Grouping — pairs which are effectively decided by the same rule are grouped together. There are thus four groups, one for each depth of rule (the last contains only one pair). The hypothesis is that this chunking is meaningful in terms of the RT distribution. This will be tested by comparing the four-group chunking of the observations with a ten-group chunking, where each group consists of the observations for a single pair. If the ten-group chunking accounts for significantly more of the variance than the four-group chunking then the latter is inappropriate. Similarly, the chunking suggested by the *ordinal distance effect* can also be evaluated. Here, the four groups consist of the four ordinal separations possible in the five-term series. For each individual, the chunking according to depth should be better (account for more variance) than the chunking according to separation.

- Linearity — if the grouping by depth turns out to be valid, then the next question is whether RT increases linearly with depth. This is to be expected if the rules are applied in a strict temporal sequence down the stack. If some

other control structure is used (perhaps a more parallel one) then there may may be some kind of non-linear, or even non-monotonic, relationship between depth and RT. The linearity hypothesis will be tested by seeing if the four-group chunking (above) accounts for significantly more of the variance than a linear regression on the same data. A similar analysis can be carried out for the ordinal distance effect with the chunking according to separation.

A comprehensive analysis of variance was thus carried out for five subjects, with each subject's data organised according to the rule stack selected on the basis of the choice data (previous section) and, in the case of White and Green, a preliminary analysis of RTs[3]. Each analysis consists of five parts, each involving all 100 observations:

1. A one way analysis of variance with ten groups corresponding to the ten pairs. This provides the base-line against which other models are compared. The ten way grouping can itself be regarded as a model — one in which the mean RTs for pairs can have an arbitrary relationship with each other. It is the most powerful model (from the point of view of fitting data), but is the least concise description.

2. A one way analysis of variance with four groups corresponding to the four depths.

3. A linear regression on the RTs, with the independent variable taking four values (1,2,3 & 4) corresponding to the four depths.

4. A one way analysis of variance with four groups corresponding to the four ordinal separations.

---

[3]The preliminary choice of stacks for White and Green was corroborated by the choice data analysis, so it is not thought that this departure from the main sequence of analyses affects the logic of the argument or validity of the statistical tests described in this section.

5. A linear regression on the RTs, with the independent variable taking four values (1,2,3 & 4) corresponding to the four ordinal separations.

The results of these analyses are organsised into two sets of tables[4] (5–26 to 5–30). One set contains analyses for the stack model and the other set is for the ordinal separation model. Each table has the same structure, and is based on four sums of squares (S.S., the measure of variance). These sums of squares were obtained from three types of computer analyses[5], a regression, a one-way with four groups and a one-way with ten groups. Note that these computer analyses were simply used to calculate the sums of squares needed for the analysis of variance tables, and were not directly used to compute significance levels themselves. As the ten group analysis us used as the baseline for interpreting the other measures (as described above), the residual sums of squares (error) from it are employed. The residuals from the regression and the four group analysis are not needed. Thus the four SSs employed in each anova table are as follows:

1. The sums of squares from the regression (Reg). This appears at the left of first main row.

2. The sums of square from the four-group anova (Four). This appears separately at the top of each table.

3. The sums of square from the ten pair anova (Ten). This also appears at the top of each table.

4. The residual sums of squares (error) from the ten pair anova (Residual). This is used as the denominator for computing all the F ratios, and appears at the left of the bottom row.

---

[4]Acknowledgement — this form of analysis was suggested to me by Francis Provan, consultant statistician for Edinburgh University Computing Services.

[5]British Medical Diagnostics Packages P1R and P1V.

The actual values used in the statistical tests are as follows:

1. The sums of squares from the regression (above). If this is large then there is a linear trend to the data (RT increases with depth/ordinal separation). If it is low, then there is no evidence of a correlation.

2. The difference between the sums of squares from the four-group anova and the regression (Four – Reg). This gives the variance accounted for by the four group model which is not accounted for by the regression. If this quantity is not significantly large then the linear model adequately describes the relationship between the four groups. If it is large, then there are deviations from the linear.

3. The difference between the sums of squares from the ten pair anova and the four group anova (Ten – Four). This gives the size of variance accounted for by the ten group chunking which is not accounted for by the four group model. If this quantity is not significantly large then the chunking into four groups is valid.

The actual tests of significance are carried out by computing the mean squares (dividing by the number of degrees of freedom) and the F ratio (dividing by the residual mean squares). The F ratios are then looked up in standard statistical tables which give the threshold values necessary for significance at the 5% and 1% levels (high values are significant). The 5% threshold values are given in the 'significance' column in the tables in parenthesese. F ratios not reaching this 5% threshold are deemed to be non-significant (N.S.).

WHITE: STACK 4					
Four-group: SS = 2030822 (3 df)			Ten-group: SS = 2970759 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	2210236	1	2210236	18.38	$p \ll 0.01$ (3.9)
Four – Reg	5046	2	2523	0.02	N.S. (3.1)
Ten – Four	755477	6	12593	1.05	N.S. (2.2)
Residual	10825593	90	20284	—	—

Interpretation — the ten-way grouping by pair is not significantly better than the four-way grouping by depth which, in turn, is not significantly better than the regression. The linear trend is highly significant with a huge F ratio, whilst the other F ratios are extremely low. Stack 4 and the depth effect provides a very good characterisation of this data. CF similar outcome for Brown.

WHITE: ORDINAL DISTANCE					
Four-group: SS = 2030822 (3 df)			Ten-group: SS = 2970759 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	1994850	1	1994850	16.58	$p \ll 0.01$ (3.9)
Four – Reg	33927	2	17986	0.15	N.S. (3.1)
Ten – Four	939937	6	156656	1.30	N.S. (2.2)
Residual	10825593	90	20284	—	—

Interpretation — The pattern is much the same as above with a slightly lower (but still highly significant) F ratio for the linear regression and slightly higher (but non-significant) values for the others. The ordinal distance effect is also a very good characterisation of this data. CF similar outcome for Brown.

Table 5–26: Analysis of RT variance for White

BROWN: STACK 8					
Four-group: SS = 618696 (3 df)			Ten-group: SS = 1112860 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	581235	1	581235	7.57	$p < 0.01$ (3.9)
Four – Reg	37461	2	18731	0.24	N.S. (3.1)
Ten – Four	494164	6	82361	1.07	N.S. (2.2)
Residual	6911256	90	76792	—	—

Interpretation — the ten-way grouping by pair is not significantly better than the four-way grouping by depth which, in turn, is not significantly better than the regression. The linear trend is significant at the 1% level, whilst the other tests are far from reaching 5% significance. Stack 8 and the depth effect provides a very good characterisation of this data. CF similar outcome for White.

BROWN: ORDINAL DISTANCE					
Four-group: SS = 553539 (3 df)			Ten-group: SS = 1112860 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	471289	1	471289	6.14	$p < 0.05$ (3.9)
Four – Reg	82250	2	41125	0.54	N.S. (3.1)
Ten – Four	559321	6	93220	1.21	N.S. (2.2)
Residual	6911256	90	76792	—	—

Interpretation — The pattern is much the same as above but with a less significantly linear trend. The ordinal distance effect is also a reasonable characterisation of this data. CF similar outcome for White.

Table 5–27: Analysis of RT variance for Brown

BLUE: STACK 3					
Four-group: SS = 5080804 (3 df)			Ten-group: SS = 8033850 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	642089	1	642089	2.30	N.S. (3.9)
Four – Reg	4438715	2	2219357	7.94	$p < 0.01$ (3.1)
Ten – Four	2953046	6	492174	1.76	N.S. (2.2)
Residual	25160580	90	279562	—	—

Interpretation — the ten-way grouping by pair is not significantly better than the four-way grouping by depth, but the regression is not significant. Depth in stack 3 provides a reasonable chunking but RT does not increase linearly with depth.

BLUE: ORDINAL DISTANCE					
Four-group: SS = 1766300 (3 df)			Ten-group: SS = 8033850 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	8682164	1	8682164	6.02	$p < 0.05$ (3.9)
Four – Reg	84136	2	42068	0.15	N.S. (3.1)
Ten – Four	6267550	6	1044592	3.74	$p < 0.01$ (2.2)
Residual	25160580	90	279562	—	—

Interpretation — The grouping by ordinal separation does not account for significantly more variance than the full regression, which is significant at the 5% level. However, the ten-way grouping by pair is significantly better than either of these. So, although there does appear to be a linear trend, the chunking by ordinal separation is not the best, leaving much variance unaccounted for. CF corresponding analysis for Green and Roger.

Table 5–28: Analysis of RT variance for Blue

GREEN: STACK 4					
Four-group: SS = 8604915 (3 df)			Ten-group: SS = 10129527 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	3555046	1	3555046	15.03	$p \ll 0.01$ (3.9)
Four – Reg	5049869	2	2524935	10.68	$p < 0.01$ (3.1)
Ten – Four	1524612	6	254102	1.07	N.S. (2.2)
Residual	21285837	90	236509	—	—

Interpretation — the ten-way grouping by pair is not significantly better than the grouping by depth. Although there is a highly significant linear trend, the four-way chunking is still significantly better. Depth in stack 4 provides a very good chunking but there are significant deviations from the linear trend. CF corresponding analysis for Roger.

GREEN: ORDINAL DISTANCE					
Four-group: SS = 284441 (3 df)			Ten-group: SS = 10129527 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	2189533	1	2189533	9.26	$p < 0.01$ (3.9)
Four – Reg	94908	2	47454	0.20	N.S. (3.1)
Ten – Four	7845086	6	1307514	5.35	$p < 0.01$ (2.2)
Residual	21285837	90	236509	—	—

Interpretation — The chunking by ordinal separation does not account for significantly more variance than the regression, which is significant at the 1% level. However, the ten-way grouping by pair is significantly better than either of these. So, although there does appear to be a linear trend, the chunking by ordinal separation leaves much variance unaccounted for. CF corresponding analysis for Blue and Roger.

Table 5–29: Analysis of RT variance for Green

ROGER: STACK 3					
Four-group: SS = 13577994 (3 df)			Ten-group: SS = 15212727 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	8705905	1	8705905	38.50	$p \ll 0.01$ (3.9)
Four – Reg	4872089	2	2436044	10.77	$p < 0.01$ (3.1)
Ten – Four	1634733	6	272456	1.20	N.S. (2.2)
Residual	20351700	90	226130	—	—

Interpretation — the ten-way grouping by pair is not significantly better than the grouping by depth. Although there is a highly significant linear trend (the highest F ratio found out of all the analyses) the four-way chunking is still significantly better. Depth in stack 3 provides a very good chunking but there are significant deviations from the linear trend.

ROGER: ORDINAL DISTANCE					
Four-group: SS = 4152520 (3 df)			Ten-group: SS = 15212727 (9 df)		
Source	Sums of Sqs	df	Mean Sqs	F ratio	Significance
Regression	3545715	1	3545715	15.68	$p \ll 0.01$ (3.9)
Four – Reg	606805	2	303403	1.34	N.S. (3.1)
Ten – Four	11060207	6	1843368	8.15	$p \ll 0.01$ (2.2)
Residual	20351700	90	226130	—	—

Interpretation — The analysis shows a similar pattern as the corresponding one for Blue. The chunking by ordinal separation does not account for significantly more variance than the full regression, which is highly significant. However, the ten-way grouping by pair is highly significantly better than either of these. So, although there does appear to be a linear trend, the chunking by ordinal separation leaves much variance unaccounted for. CF corresponding analysis for Blue and Green.

Table 5–30: Analysis of RT variance for Roger

5.4.1 Summary and Conclusions

Only two subjects, White and Brown (stacks 4 and 8), appear to provide unambiguous support for the stack model. For both these subjects there appear to be linear depth effects which account for more variance than the ordinal distance model (which is a descriptive, not a computational model). For the three other subjects, the *grouping* hypothesis appears to be correct but there is not a straightforward linear relationship between the RTs for the four depths. For one subject, Blue, there appears to be no linear trend at all, whilst Green and Roger appear to show zig-zag like deviations from the trend. The ordinal distance model also fares badly with respect to Blue, Green and Roger. Although there is a significant linear trend in each case, it appears that the data do not chunk naturally into the four ordinal separations. Where ordinal distance *is* a good description (for White and Brown), the depth effect provides an as-good or better characterisation.

The reaction times for the subjects are plotted against depth of rule in figures 5–1 to 5–5 at the end of this chapter. The RT for each pair is plotted separately so that the clustering of RTs at each depth can be observed. Broken lines indicate the line found by regression, where this is significant, and solid lines indicate the relationships between the mean RTs at each depth, where this is the best description. The RTs for Bump are also plotted for comparison, although the linear regression was only carried out on the means for each pair (10 data points). Nevertheless, the trend is significant ($r = 0.86$, $p < 0.01$, t-test, 9df). The slope of the linear trends which were significant were as follows, White 149 mS/Rule, Green 189 mS/Rule, Roger 295 mS/Rule, and Brown 76 mS/Rule. If the model were correct, it would appear that there is a fair degree of variation in the time it takes the subject to test the precondition of each rule and to move on to the next — from about 80 mS to about 300 mS, almost a factor of four.

5.4.1.1 Conclusion

It seems as though the stack model captures the logic of subjects strategies but not the details of actual processing (at least, not in all cases). That subjects behave as though they employ a stack of rules as a logical sequence, is supported by the fact that the reaction times for the pairs appear to chunk naturally according to depth of rule for every subject. In other words, pairs which are decided by the same rule take the same time to process. However, the simple idea that the rules are applied in *temporal* sequence (down the stack) appears to be at least partially incorrect. Three subjects, Blue, Green and Roger show a non-linear relationship between depth of rule and RT (although significant linear trends exist for Green and Roger). If these deviations from the linear are genuine phenomena, as the evidence strongly suggests, then this means that subjects' representations must be considerably more complex than previously supposed. This detracts somewhat from the attractive simplicity of the stack model, but perhaps it is not too surprising that this turns out to be necessary. In particular, it seems that it will no longer be possible to think of the ordering of rules as being a purely temporal/procedural one; subjects must somehow explicitly represent the order of the rules independently of their application. Also, I would have expected the time taken to step through the rules to be more consistent between subjects if this was the primitive mechanism, though this is just an intuition.

Although the above analysis shows that the proposed 'depth effect' is not a unitary phenomenon (but just a trend in the data), it also shows that the ordinal distance effect is similarly artifactual, in that the reaction times do not chunk naturally into ordinal separation categories. What is clear is that there are strong regularities in the data which are incompletely captured by both models. It seems likely that the variation in RT is caused by a mechanism for invoking and applying avoidance and selection rules, but that the mechanism is more sophisticated than a simple run down a stack. Other possible sources of RT variation which were entertained were:

1. avoidance rules taking longer.

2. differences between training and remote pairs.

3. variation with the number of rules applicable to a pair.

Whilst these possibilities have not been definitely ruled out, no systematic variation in RT with these factors has been detected.

5.5 Overall Conclusions

The stack model may be incorrect in subtle ways but, overall, it is a remarkably simple and effective approximation of monkeys' decision procedures during the five-term series task. The data from seven subjects and a number of experiments has been analysed at several different levels. With the possible exception of one subject (Blue), all the analyses lend support to the model.

In more detail, the stack model appears to be correct in the following aspects:

1. The distinction between *alpha* and *random* types of triads appears to have a great deal of validity, and supports the notion of subjects employing a combination of avoidance and selection rules. This is also supported by the clustering of RTs of pairs which would be discriminated by the same rule.

2. There appears to be some mechanism by which rules are ordered, but this may not be a rigid stack as originally supposed. This is supported by the ability of the model to account for much of the variation in choice patterns between triads and, to a lesser extent, the overall linear trend in the RT data for all subjects except Blue.

3. The improvement in performance observed across the three triadic testing phases can be partly accounted for by increasing iteration in the control strategy for applying the rules. The residual improvement can potentially be accounted for by the stack becoming more 'rigid' (firmly ordered) with usage, although there is no formal model for this.

There are a number of problems with the model however, some of which are summarised below.

1. The model predicts all the choices to go to the α item in *alpha* type triads, whereas a proportion get diverted to β. It seems likely that this is because the assumption about rule order being fixed is partially wrong.

2. As discussed in the last section, there is mismatch between the linear binary RT profiles predicted by the stack and those obtained, particularly in the case of Blue, Green and Roger. However, the ordinal distance effect does not characterise these subjects well either.

5.5.1 Further Work

Triadic RTs

Working out the implications of the partially iterating control strategy for RTs is not simple as there are many parameters to specify. The implications of the *naive* strategy are simpler, but unfortunately RTs were not taken in the *early* phase where this strategy is a reasonable approximation (according to the choice data). With an iterating strategy, the problem is that rules get used in combinations so there is no simple 'depth' prediction. Furthermore, we can no longer assume a uniform time for the testing of preconditions of rules because of the elimination of items. Initially, there are three items to check for relevance to each rule but this is reduced to two after deciding to avoid one item.

Another question is what happens on triads where there is partial iteration? Should trials on which there is a lack of iteration be faster? Perhaps iteration is attempted every trial but sometimes aborts. It may be possible to answer this question empirically. We already know that, overall, triadic RTs increase with improved performance, but it remains to be seen whether this is selective between trials and triads.

Other sources of data

It has been observed (informally) from the video tapes of the monkeys performing the triadic task that some subjects will often try to pick the second best choice after they have correctly picked the α item. They are prevented from doing this by the termination of the trial by the experimenter (the objects are withdrawn from view), but it is often clear which item would have been the subject's second choice. This is not predicted by a simple stack driven model in which the application of a *selection* rule would leave the subject with no means of determining the 'second best' (β) item. This is because the time between the subject making its first and (would be) second choice is, almost certainly, too short for the decision procedure to be invoked again for the remaining pair. This suggests that, at least in some circumstances, that at the time of making the decision, the subject has more than one relevant rule available for governing choice.

The process of reducing the data inherent in the video tapes into a manageable form (as formal protocols) is beginning to be undertaken. It may be possible to use such measures as the time and order each item is attended to to disambiguate between possible decision procedures.

Finally, there is data from children which might fruitfully be analysed according to the stack model. Children have been tested under analogous conditions to the monkeys (including triadic tests). Although there is possibly not enough data to model individuals in detail, it ought to at least be possible to test the averaged stack model.

Suggested Experiments

The stack model could be quite simply tested if it were possible to test subjects with pairs containing a 'neutral' item which had previously neither been 'rewarded' nor 'punished' with respect to the other items, and yet was familiar enough not to cause an aversive 'novelty' reaction. Such pairs could be used as 'probes' to discover which (if any) stack the subject was using. For example, a

sixth colour in the five-term series experiment would sometimes be chosen and sometimes not, depending on which of the five trained items it was paired with. Such an experiment would be possible with Von Feren's paradigm (chapter 3), where absense of a reward is not synomymous with an incorrect response. During training, it would be possible to have the neutral items ocasionally paired with the main ones without giving either positive or negative feedback. This would hopefully familiarise the subjects with the neutral items without them becoming incorporated into the series.

Another possible extension to the experiment would be to test subjects with quadruplets in addition to pairs and triads (triplets). There are five possible quadruplets in a five-term series, each with one of the five items missing. Quadruplets would be most interesting for subjects which appeared to be employing more than one *avoidance* rule. In some such cases, three rules would need to be applied in order for a unique choice to be made. For example, if Brown, White or Green were to be given the items A, B, C, and D to choose between, then the successive application of the rules:

$$present(A) \Rightarrow avoid(A)$$
$$present(B) \Rightarrow avoid(B)$$

— would not be sufficient to allow a unique choice between C and D. If a subject could cope with triplets but made random choices in situations such as the one above, this would be indicative of some kind of resource limitation, such that there was a limit on how many rules could be applied. If, in such case, a subject performed better with a quadruplet than a triad, this would be evidence against the stack model.

Some further suggestions for experiments are made in chapter 8.

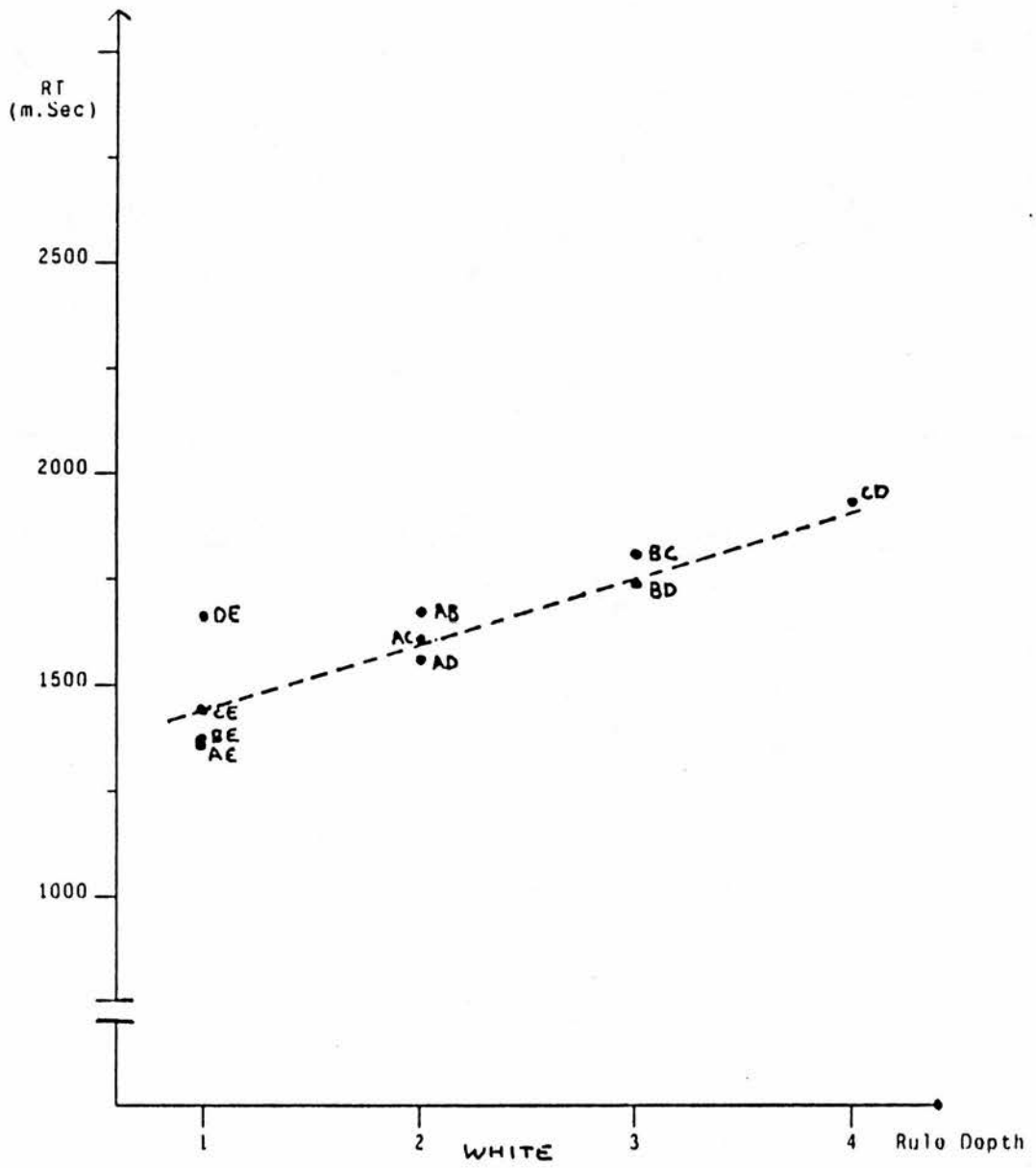


Figure 5–1: White's binary RTs plotted against depth of rule in stack 4.

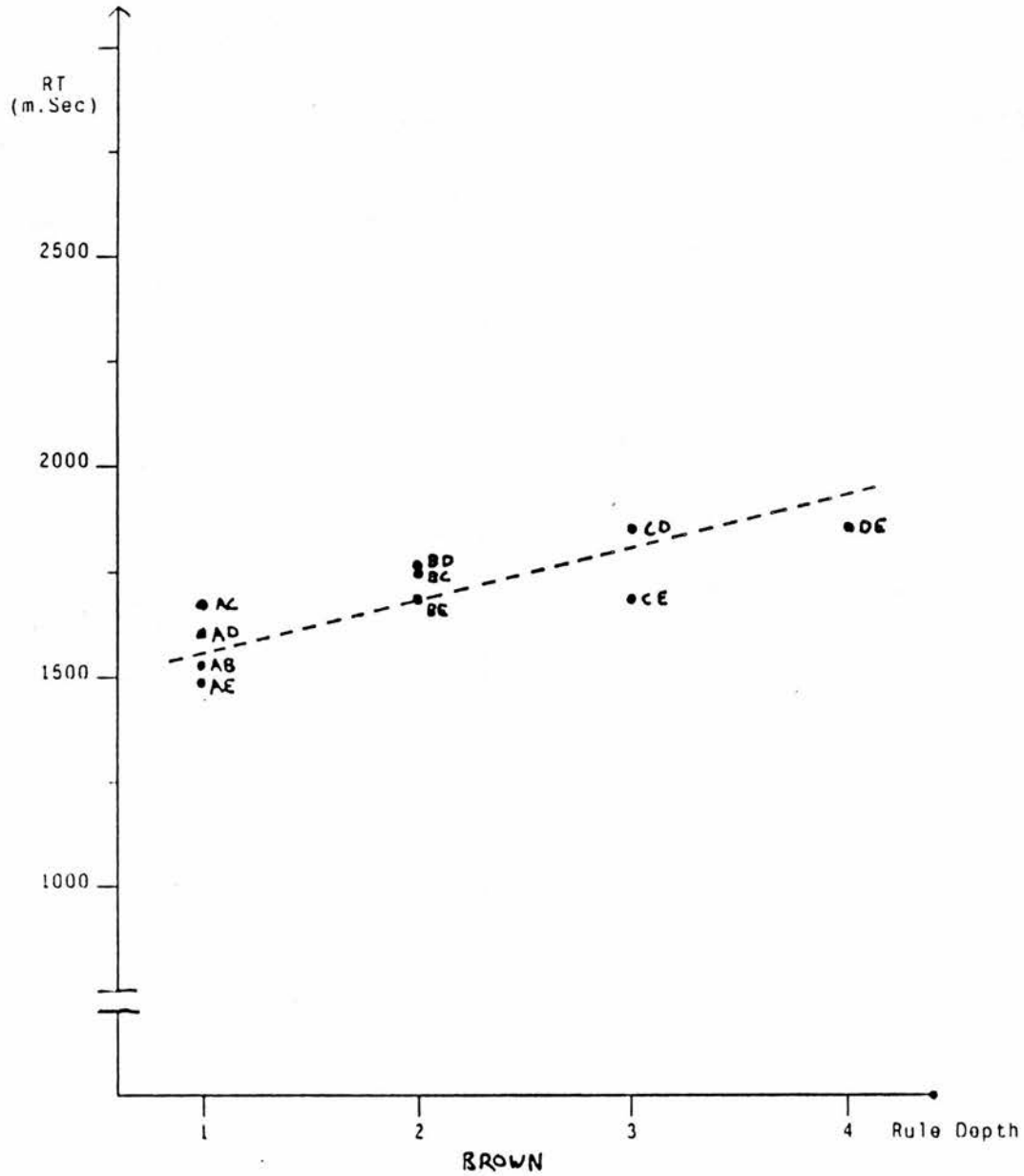


Figure 5–2: Brown's binary RTs s plotted against depth of rule in stack 8.

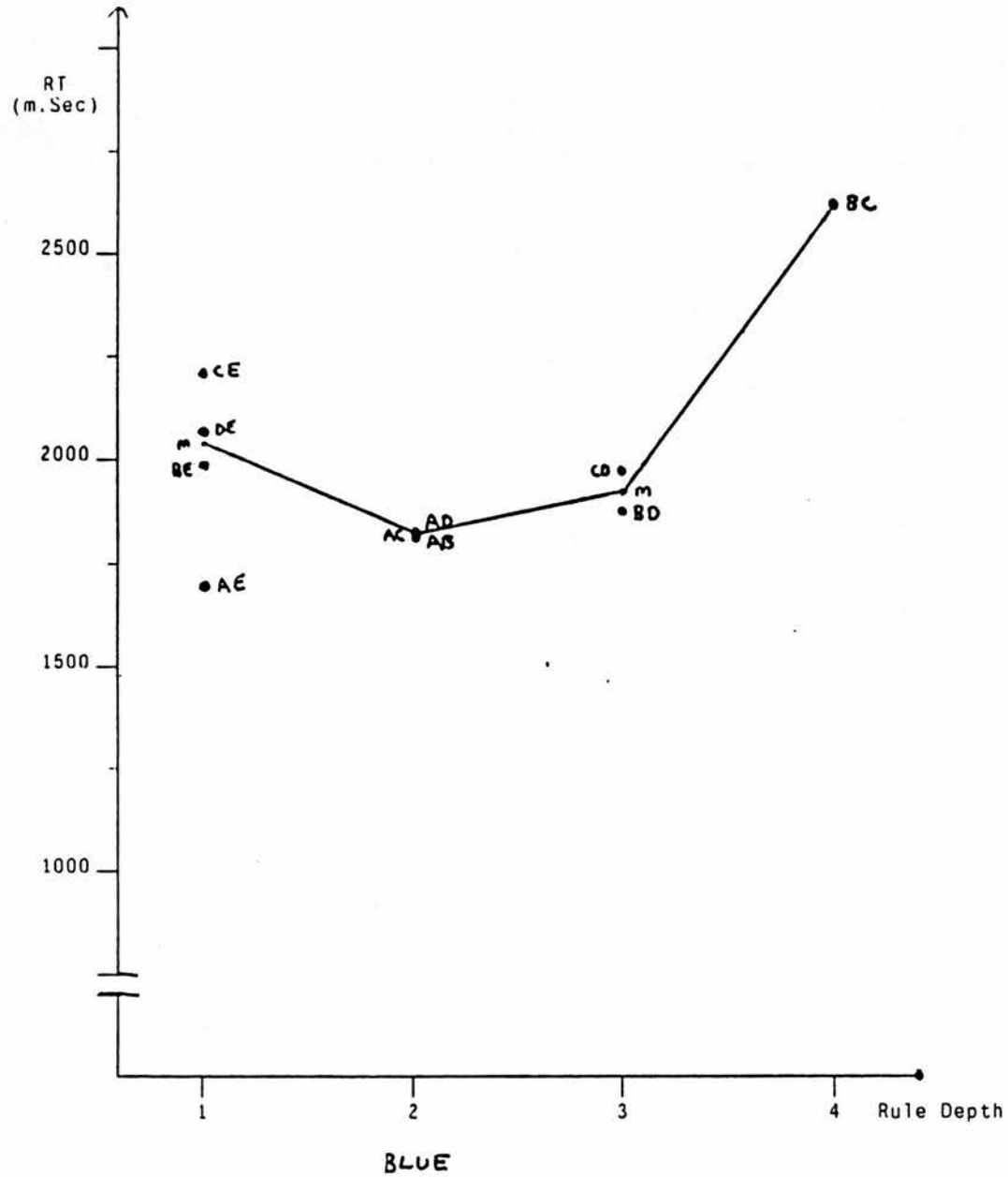


Figure 5–3: Blue's binary RTs plotted against depth of rule in stack 3.

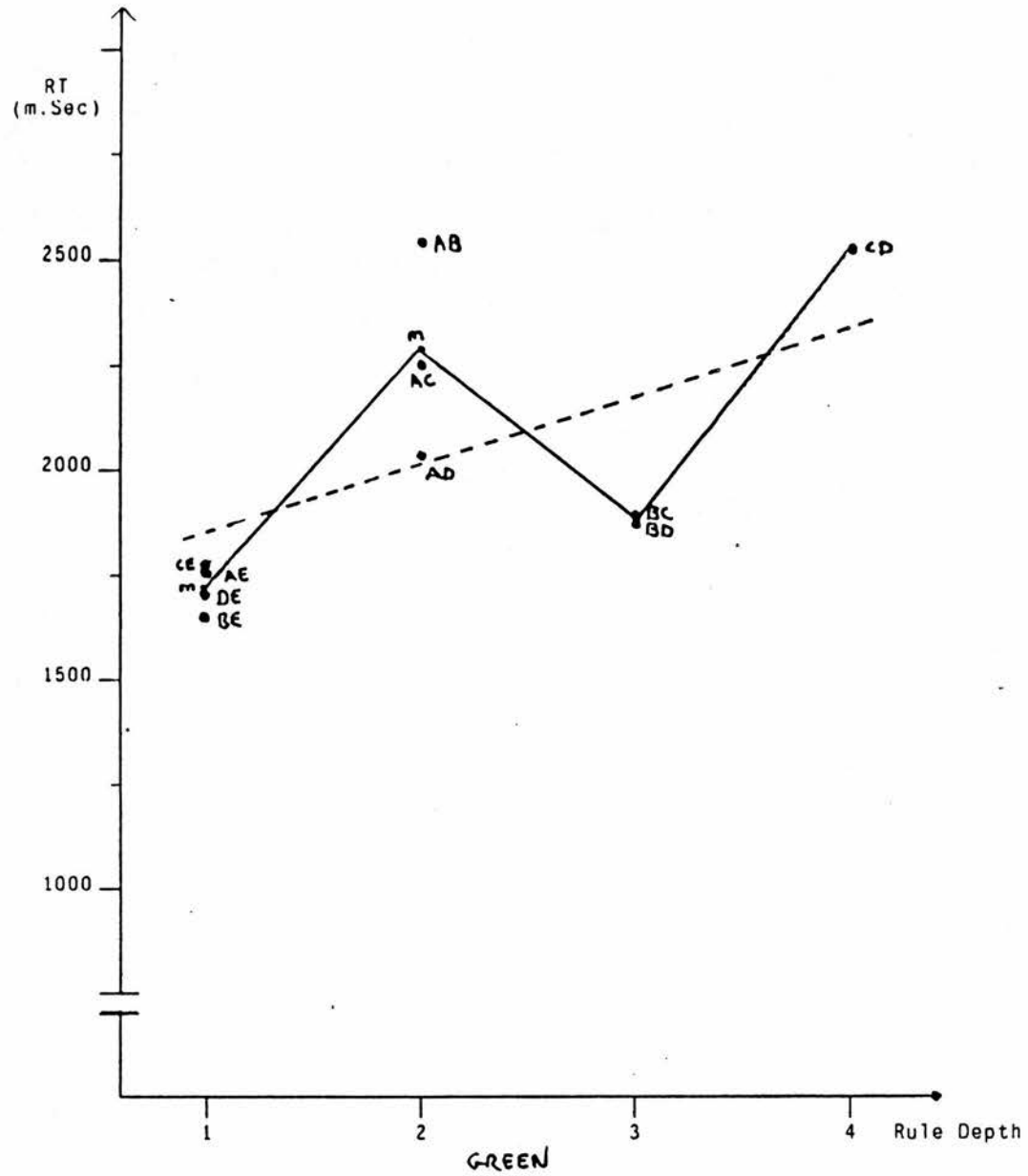


Figure 5–4: Green's binary RTs plotted against depth of rule in stack 4.

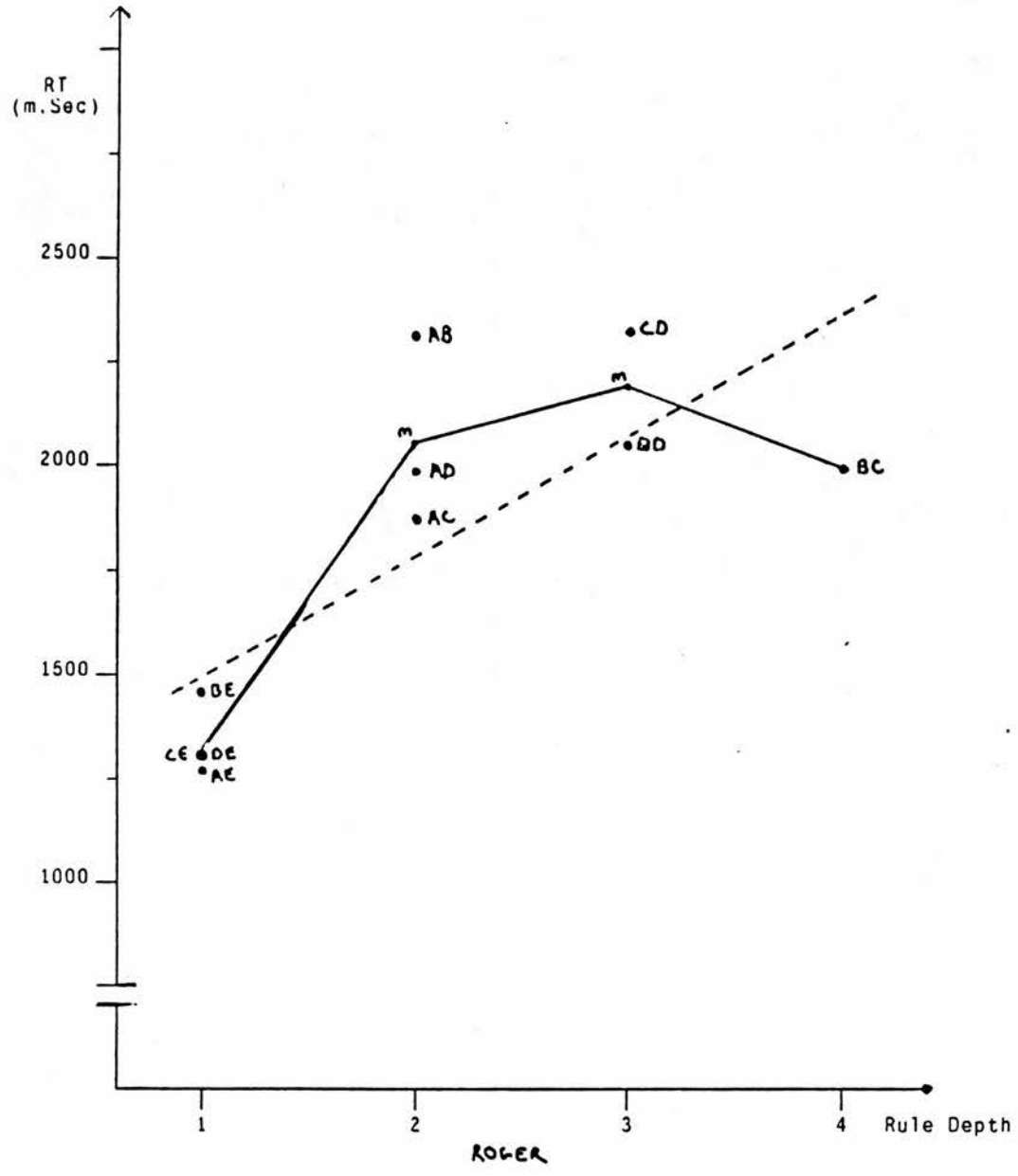


Figure 5–5: Roger's binary RTs plotted against depth of rule in stack 3.

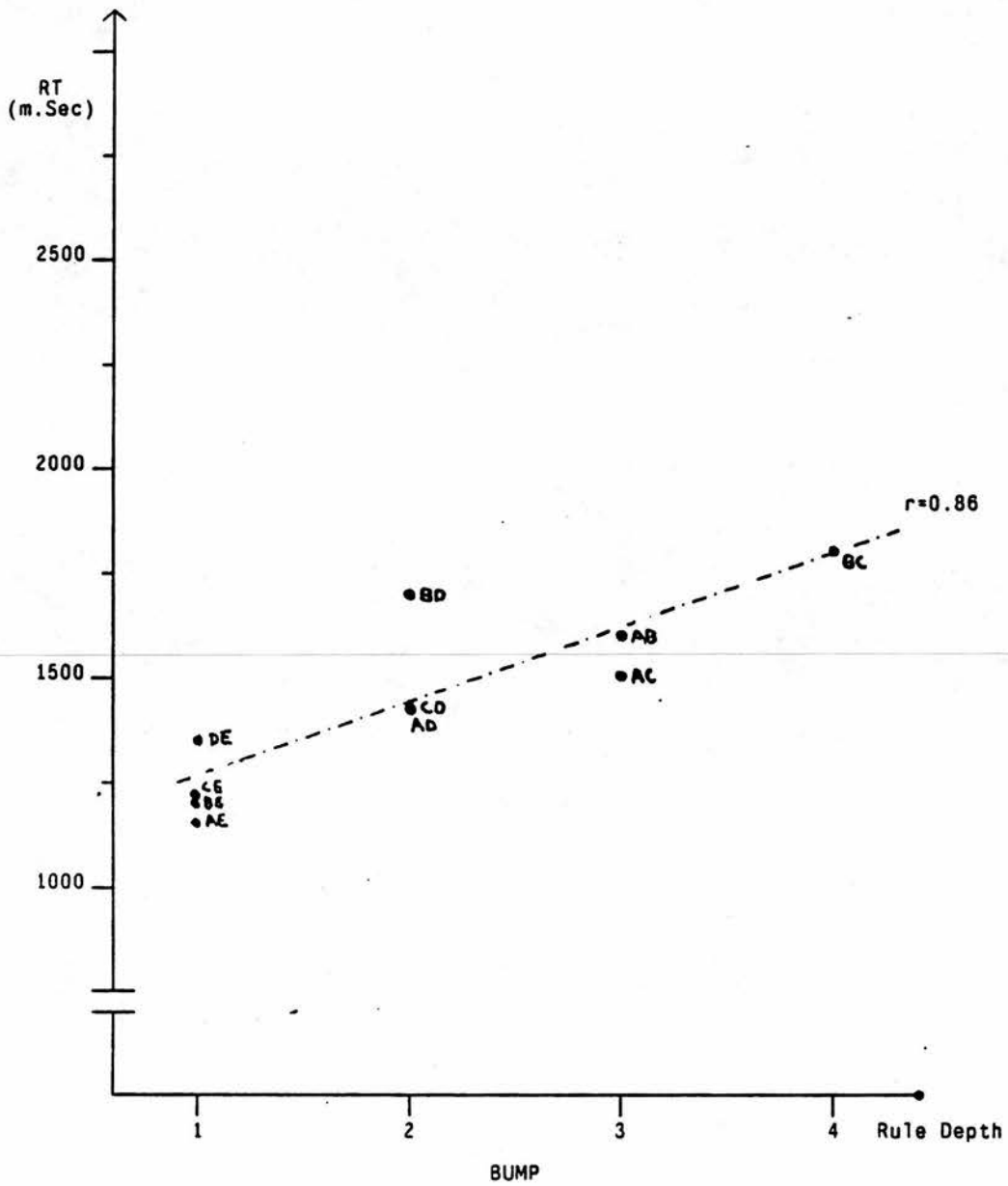


Figure 5–6: Bump's 1978 binary RTs plotted against depth of rule in stack 2.

Chapter 6

Deductive Processes in Transitive Inference

Having presented an account of how monkeys solve the N-term series problem, this chapter goes into a deeper analysis of existing models of transitive inference, introduced in chapter 2, and compares them with the stack model. It is argued that most rely, at some level, on hypothesising the existence of a transitive inference rule and deductive reasoning. The comparison includes a re-examination of Trabasso and Breslow's models (introduced in chapter 2) and various AI approaches to transitive inference and the SDE phenomenon. Both the quantitative predictions and the theoretical properties of models are considered.

6.1 Trabasso's and Related Models of the Distance Effect

Experimental investigation into human performance of transitive inference started as a study of abstract deductive reasoning. Formal deduction was, in the time of Burt and Piaget and right up until recent years, the major model of human reasoning (Gregory, 1981; Johnson-Laird, 1983), or at least, the major paradigm for thinking about human reasoning. Johnson-Laird refers to this as the 'doctrine of mental logic'. Thus even when Trabasso discovered that children performed the task by constructing some kind of integrated representation of the starting premises, he still assumed that this was a deductive process involving a transitive inference rule. The inference problem got pushed one level deeper and

the investigation concentrated on the retrieval aspects. Only Breslow tried to give a detailed account of how pairwise information might be combined without appealing to an innate transitive inference schema. However, it turns out that his explanation is only superficially different from Trabasso's on this count, as is demonstrated below.

Trabasso and Riley (Trabasso & Riley, 1975) argued that the end-anchor, marking and distance effects could only be explained by hypothesising that subjects employ transitive inferences to construct a linear order (seriate the objects) at the time of learning the premises. The distance effect is then due to the 'look-up' processes for locating or comparing the positions of two objects within the order to find which comes first and which second. The basic idea is that the further points are apart in the linear representation, the easier it is to discriminate them. Although Trabasso does not explicitly say it, this is presumably supposed to be analogous to perceptual discrimination.

6.1.1 The Training Phase

Although Trabasso *et al*'s experiments and models are mostly concerned with the testing phase, they do say that the series is constructed in an 'end-inwards' fashion, reflecting the finding that subjects learn the premises involving end terms first and learn premises involving middle terms last. There is no detailed computational account of how the series is constructed but ordered pairs are supposed to be combined transitively;

$$< A\ B > \&\ < B\ C > \longrightarrow < A\ B\ C >$$

so that some kind of vector is formed by forward inference. This is not free standing, but is associated with a linear scale as I have tried to illustrate in figure 6–1.

It is not at all clear whether the objects are being mapped onto a pre-existing set of codes or whether the order is constructed as a function of the information being integrated. However, the net result is that there is some kind of linear

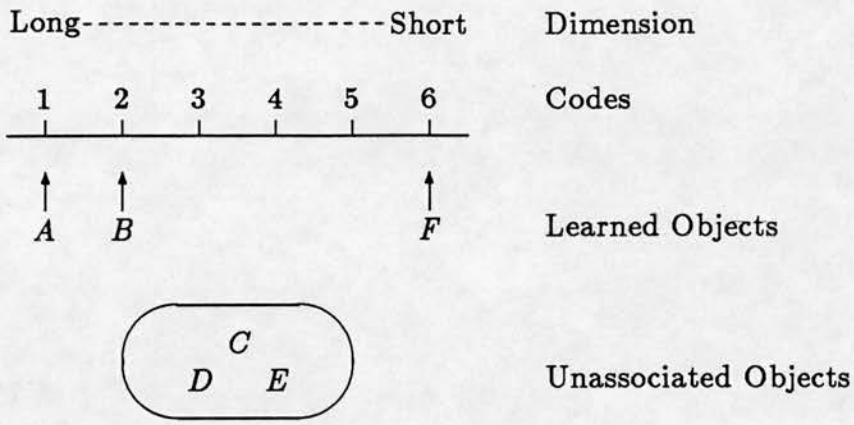


Figure 6–1: Partially learned series — Trabasso's model

scale (perhaps with a spatial metric) onto which object labels become mapped. The first objects to become associated are the end terms as these have only one comparative identified with them. Once this has happened, information from ordered pairs allows further objects to be inserted into the line. The pairs, AB and EF, allow objects B and E to be associated, followed by pairs BC, DE being assimilated, and so on. If there are errors associated with each of these steps then this helps to explain the serial position effect in training — the performance is poorer the nearer the pairs are towards the middle of the series.

6.1.2 The Testing Phase

Trabasso *et al* propose two explanations of how use of a linear order may give rise to the distance effect: the *associative strength model* and the *spatial discrimination model.*

Both give identical predictions about the rank order of reaction times for different pairs. In one model the 'confusability' of the codes for different objects is the relevant factor, while in the other, the (spatial) 'discriminability' of each object from others in the series is what affects decision times. Both models make quantitative predictions rooted in information theory, however, no specific mech-

anisms for comparison or location of objects are proposed and many arbitrary assumptions are necessary.

In the *associative strength model* each object is associated with several codes in the array with differing degrees of strength. The idea is that an object is initially associated with one code (as described above) but that somehow the association 'rubs off' onto adjacent codes, declining in a monotonic fashion away from the main point of association. This is supposed to reflect the fact that 'very short' is similar to 'short' etc. and is called 'generalization'. The initial strengths of the associations depend on a number of factors including the 'availability' of the code (reflecting the marked/unmarked difference) which gives rise to the distance effect. However, as the latter aspect of the account is somewhat vague (the model already has too many degrees of freedom), we must suppose that all initial strengths of association are equal. The idea of variable strengths of association is akin to the idea of variable rule priorities, as discussed in section 4.6.1.

The way objects are compared, as described in (Trabasso & Riley, 1975), is to compare their relative strengths of association (after the process of generalisation has taken place). Here, the 'relative strength' is the ratio:

$$RS = \frac{\text{strength (of association) from an object to a code}}{\text{summed strengths of object to all codes}}$$

This is supposed to give the probability of association of an object with a particular code, but it is not made clear at what point this ratio is computed, if at all, by the subject. Perhaps it should be regarded as a mathematical model of the *end result* of generalisation, rather than having any bearing on the process itself. It is then assumed that 'the ease of generating a pair of codes is directly related to the sum of the code probabilities' and it is these summed probabilities that give the reaction time predictions for each pair. It is not clear why this should be so.

Note that the success of this prediction algorithm depends on the fact that comparisons are not local but depend on codes for the entire series. If the relative discriminability had to be directly calculated by the subject this would

seem to imply a computational commitment way beyond the complexity the task demands. However, Trabasso *et al* are not really proposing a mechanism — they are giving a quantitative mathematical description of the behaviour of some hypothetical decision making machinery. In the same way, information theory does not prescribe choice algorithms — it describes 'virtual' information processors, for example: what can be maximally done with a noisy set of data. The authors employ notions of probability and they are not clear whether they are proposing a stochastic mechanism which will show the distance effect over a large number of trials or whether the 'confusability' of codes somehow leads to reaction times proportional to the uncertainty. The problems mentioned above also apply to the second model, described below.

The *spatial discrimination model* attributes the differences in reaction times to differences in the ease of locating objects within a spatial array. Again, no specific mechanism is proposed but a metric of 'relative discriminability' is given:

$$\text{Relative Discriminability} = \frac{\text{sum of distances from object to other objects}}{\text{sum of all the distances between objects}}$$

This measure is higher for items near the ends of the array, reflecting the idea that these terms are somehow easier to find. As both objects have to be located before they can be compared, the sum of their relative discriminabilities is hypothesised to be inversely proportional to the reaction time for that pair. This produces a predicted scale of reaction times exactly the same as for the more complex association model above, and so the simpler model will be assessed.

For the five-term series, for example, the sum of all the distances between items is 10, assuming a unit distance of 1. The total distance of the item B from the other four items is $1+1+2+3=7$, giving rise to a relative discriminability (RS) of 0.7. The RS of C is 0.6, so the predicted RT for the comparison BC is:

$$c + \frac{k}{0.6 + 0.7}$$

— where c and k are constants. Table 6–2 shows the predictions for each pair scaled so that the full range of variation is 0 to 100. The same scaling mechanism

has been applied to the monkey RTs for ease of comparison. The monkey data are binary RTs averaged across the subjects which were analysed individually in section 5.4, excluding Bill, who was tested at a different time. The rest of the table is explained further on.

Taking the overall ranking of projection times, the projection from this model is not bad. It correctly projects higher RTs in the middle of the diagonals ($\nwarrow$)[1] and, of course, decreasing RTs with ordinal separation ($\nearrow$ towards AE). However, Trabasso's model incorrectly predicts some local exceptions to the distance effect, for example AB being faster than AC. Also, it does not predict the overall asymmetry in the monkey data and predicts too low an RT for the pair BD.

The averaged stack projection was generated by combining the projections from all eight stacks for each pair, and is also symmetrical. An asymmetrical projection would result if there was an overall bias in favour of *selection* over *avoidance* rules in the stacks, as was the case in the analysis of individuals. The average of all eight stacks is given here for generality, and it can be seen that it fares better than Trabasso's model on all the points raised above. Furthermore, Trabasso's model has no principled way of accounting for individual variation and makes no predictions about error rates on the triadic tests.

[1]The arrows refer to directions in the table.

6.2 Breslow's Sequential Contiguity Model

The model about to be described was invented by Breslow as part of an argument against Trabasso's claim to have found coordination of transitive relations in children. Here we examine it in its own right for its ability to explain Trabasso *et al*'s empirical data on the five and six term series tasks.

Like Trabasso's models, this consists of two parts: information from separate premises is integrated into a unitary representation during the acquisition phase, and there is a procedure for interrogating this representation in order to answer questions during the testing phase. As Breslow also supposes that subjects form a linear ordering, the two parts of his model can be assessed separately.

The main idea in the learning phase is that an understanding of the premises as order relations is not a necessary precondition for constructing a linear order. Young subjects treat the premises as categorical statements about the absolute properties of objects (as has also been suggested by other authors). 'X is longer than Y' is interpreted as 'X is long' and 'Y is short' and also that X 'goes with' Y. Both comparatives are given to the subjects so the information initially appears contradictory with all but the end terms appearing sometimes long and sometimes short. However, the contiguity between pairs that are always mentioned together will also be learned. Although they do not start out to form a linear order, it supposedly materialises out of contiguity relationships, growing end inwards from the unambiguously labelled long and short end items.

Breslow argues that this does not involve transitive inferences:

> 'A becomes related to B, which becomes related to C, and so on, without the generation of any higher order relation such as "A is related to C". Further, the fact that they learn to produce the linear order from both ends does not imply that they deduce the reversability of the ordering. For instance, they do not infer the ordering from E to A from the already established ordering from A to E. Rather, they simply learn the linear ordering in each direction separately and in the same nonreversible fashion.'

Long: $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f$

Short: $f \rightarrow e \rightarrow d \rightarrow c \rightarrow b \rightarrow a$

Figure 6–2: Long and short chains — sequential contiguity model

I have attempted to illustrate this idea in figure 6–2. The problem with this argument is that the 'higher order inferences' that Breslow is calling for as evidence of transitivity can inevitably not be found in the integrative stage alone. Inferences only manifest themselves when representational and interrogative processes both happen. It is no good putting part of the information process under the microscope and asking "where's the inference?". Incidently, this is also an argument against the opposite theoretical position as espoused by Johnson-Laird (Johnson-Laird, 1983), who claims that all deductive inference is mediated without (logical) inferential machinery. However, this debate about whether information processes are 'logical' or not is something of a diversion. Let us continue with the second part of Breslow's model — that for question answering.

Subjects make comparisons by assessing the linear order in the same end-inwards sequence in which they originally constructed it. If the comparative in the question is 'longer' , as in 'which is longer, X or Y?', then they start from the 'long' end of the sequence until they come across either X or Y. Whichever comes first is deemed to be the 'long' object, in categorical fashion, as it is nearest the 'long' end of the ordering. Again, Breslow states that this process requires no transitive inferences, although this is a matter of perspective.

If it is assumed that each traversal along the sequence takes a constant amount of time, then a distance effect is produced: the further apart two ob-

Which shorter? \ Which longer?	A	B	C	D	E
A	–	0	0	0	0
B	3	–	1	1	1
C	2	2	–	2	2
D	1	1	1	–	3
E	0	0	0	0	–

Table 6–1: Traversal times for Breslow's model.

jects are, the nearer the chosen object will be to an end anchor and the fewer traversals will be needed. However, this is a very asymmetrical effect, being different for each of the two comparative forms of question. Table 6–1 shows the number of traversals needed for each comparison for both 'longer' and 'shorter' type questions. The table is for a five term series in which A is the long end anchor and E is the short one. The diagonals show comparisons between pairs which have the same ordinal separation and the means of these give the distance effect. Because the table has rotational symmetry, the distance effect for either question form is the same. The success of Breslow's model with respect to other aspects of the RT variation crucially depends on averaging RTs across the two question forms. Table 6–2 shows these averaged traversal times scaled to vary from 0 to 100, for comparison with the other models. Notice that these averaged RT projections contain no element of variation other than the distance effect.

In order to account for other aspects of the RT variation, such as the higher times for pairs towards the middle of the series, Breslow made some additional assumptions. Basically, if, when the appropriate anchor is retrieved, it turns out to be one of the comparison items, the accessing time is assumed to be shorter. Furthermore, this advantage is greater for the long anchor than for the short anchor, thus producing an asymmetry. With these adjustments, the ranking of the fifteen possible pairs closely correlates with the ranking from Trabasso's reaction time experiments. For the five term series, the projection becomes very similar to that of the stack model.

However, the stack model makes no special assumptions about end-anchors and, furthermore, it subsumes Breslow's basic model. Stacks 1 and 8 (in conjunction with a control strategy) are computationally equivalent to traversing one or the other of Breslow's chains. Averaging the projections from stack 1 and 8 produces an identical projection to that shown for Breslow's model in table 6–2. The relationship between the two models is explored further in section 6.4.

A final point about Breslow's model is that it predicts wildly varying reaction times for certain pairs depending on which comparative appears in the question. For example, the number of traversals for pair AB is three or zero depending on whether the question is 'which shorter?' or 'which longer?' This prediction cannot be tested with the monkey data and, unfortunately, Trabasso *et al*'s results for the two comparatives were only reported lumped together. The only exception was the *congruence* effect reported for end-anchor pairs — when the question matched the end anchor label, the reaction time was faster. However, it seems that this effect was a fairly small in comparison with the distance effect. It seems likely that individual variation is more important.

6.2.1 Concluding Remarks — the Trabasso vs Breslow Debate

On the one hand, we have Breslow discounting Trabasso's work as being irrelevant to true transitivity on the grounds that he can model the behaviour of Trabasso's subjects with 'non-transitive' techniques. He suggests that this is a computational strategy using low-level categorial processing which is set up in response to practice at a particular task. It is thus irrelevant to measuring attainment of a concept of transitivity or classical logical behaviour. How then are we to interpret the work on the five and six term series, Breslow's own, and related information processing models? Is it all of no psychological interest?

On the other hand Trabasso claims that this task is relevant to a variety of behaviours: the whole domain of processing information with transitive properties including the work on seriation (Inhelder & Piaget, 1958) and symbolic

Monkeys ($N = 5$)

⋱	B	C	D	E
A	84	58	55	0
B	–	75	73	15
C	–	–	100	35
D	–	–	–	27

Trabasso Mod.

⋱	B	C	D	E
A	33	46	33	0
B	–	100	59	33
C	–	–	100	46
D	–	–	–	33

Av. Stack Mod.

⋱	B	C	D	E
A	41	35	24	0
B	–	100	71	24
C	–	–	100	35
D	–	–	–	41

Breslow Mod.

⋱	B	C	D	E
A	100	67	33	0
B	–	100	67	33
C	–	–	100	67
D	–	–	–	100

The monkey data and the predictions from Trabasso's, Breslow's and the averaged stack model are all scaled so that the quickest pair is zero and the slowest is 100. Note that all the model projections are symmetrical around the diagonal axis (↗), whereas the monkey data is not. With some additional assumptions (see text) the projections of Breslow's model converge with those of the stack model.

Table 6–2: Scaled variation in RTs — monkey data compared with models

comparisons (Moyer, 1973). Yet he offers no explanation of logical competence and almost seems to be saying that the problem is the solution: although he supposes children have the ability to coordinate transitive relations much earlier than Piaget claims, and calls on information processing ideas to support his argument, he has cornered himself into the position of being a 'Nativist'[2] with all its attendant problems as discussed by Johnson-Laird (Johnson-Laird, 1983).

An alternative to these perspectives is the view that not only are transitivity tasks related to other types of inference as Trabasso states but that *the information processing analysis is important in its own right.* The study of transitivity tasks gives us a vehicle for studying representation, strategy and other topics fundamental to cognitive science. Furthermore, we may begin to be able to offer an explanation for the emergence of conceptual transitivity and logical competence in general. Breslow states that he knows of no information processing type models which explain generation of processing strategies from general cognitive structures. It seems to me that this is a good statement of what the goal of Cognitive Science should be and of what we should be working towards in the domain of transitivity tasks. The acquisition model discussed in chapter 7 may be a small step in this direction.

6.3 Discrimination Trees

Bundy wrote a Prolog program[3] in which information about the relative positions of a set of objects along an ordinal scale was stored in the form of a discrimination tree[4]. The program is included in appendix B.1.

---

[2]Believing that logical skills are innate (inherited)

[3]'Modelling McGonigle's 'Bigger-Than' Data With a Discrimination Tree', Dept AI, Mathematical Reasoning Group, internal note

[4]Also called a discrimination net.

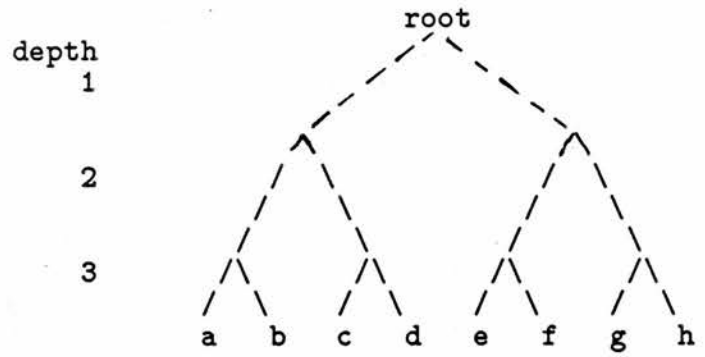


Figure 6–3: Discrimination tree representation of eight term series.

Bundy picked a discrimination tree to base his model on because it had the right 'gross' timing properties (it shows a distance effect). 'Timing' refers to the number of operations necessary to make a decision, in the same way as for the stack model. Using a discrimination tree, the objects of the series are represented as the leaves of a binary decision tree so the number of key decisions is equivalent to the depth of the search before a divergence in the representations of the objects is found. However, as we shall see, with this representation the relationship between the ordinal separation of a pair of objects and decision time is not monotonic.

For example, in figure 6–3 D and E diverge at the root whereas C and D do not diverge until a node three levels deep. Both these pairs have an ordinal separation of 1. The decision 'times' (depth) for different pairs in the series $A - H$ are tabulated in table 6–3.

This shows clearly that the time for pairs of any given ordinal separation smaller than four (with this size tree) is not constant but varies in a convoluted way with the ordinal position of the pair. As with Breslow's model, it is only when averages are taken that the times reflect the general distance effect. For example, the discrimination tree would appear to predict a particularly quick response to questions referring to the pair DE (or, in general, pairs which straddle the two main branches of the tree). This effect is not found with the monkey

	B	C	D	E	F	G	H	mean (separation)
A	3	2	2	1	1	1	1	
								\1.0 (1)
B		2	2	1	1	1	1	
								\1.0 (2)
C			3	1	1	1	1	
								\1.0 (3)
D				1	1	1	1	
								\1.0 (4)
E					3	2	2	
								\1.4 (5)
F						2	2	
								\1.7 (6)
G							3	
								\2.4 (7)

Table 6–3: Discrimination times for different pairs and means for each ordinal separation (diagonals).

subjects, even at the individual level (section 5.4). No such effect was reported by Trabasso *et al* either, although it might seem that the dividing point in the series between the two main branches varies from individual and the effect has been 'averaged out' in pooling data from many individuals.

The discrimination tree model could still be a candidate for modelling the distance effect in the SDE studies (see chapter 2). However, the 'straddling' effect has not been found to occur in the symbolic comparison task (McGonigle & Chalmers, 1984a) where the subjective midpoint of a series has been found to be remarkably stable between and within subjects. Children were quite willing to categorise animals with a range of sizes as either 'big' or 'small' and the 'cut-point' seemed to be in the middle of the series and thus dependent on the range of sizes in the task and not on any absolute metric.

A number of other researchers have suggested 'categorical' models of the distance effect in which the range of sizes, or whatever, is chunked into two or more categories. The further apart two objects are apart the more likely it is that they will fall into different categories thus enabling a simple and fast decision.

Kosslyn et al have suggested 'dual process' models in which categorical and analogue comparators operate sequentially or concurrently (Kosslyn *et al*, 1977). However, categorical comparison of this kind does not appear to occur in the transitivity task. For example, Woocher et al trained subjects on two separate arbitrary rankings and then combined them into one ranking with an additional linking premise. The serial position effects indicated that the subjects used one long series for the comparisons and were not significantly faster on comparisons between the two original series (Woocher *et al*, 1978).

For a symmetrical discrimination tree, such as the one illustrated in figure 6–3, there is no privileged end of the series and no sense in which the series is directional. The technique is neutral with respect to comparative and so there can be no interaction between serial position and comparative and so no *congruity* effect.

However, this is the closest to a simple computable model that I can think of for what Trabasso had in mind when he was talking about 'discriminability' of objects in a series. The tree representation of a series can be viewed from the root to the leaves as a successively finer grained representation of a linear space. Large separations can be resolved immediately using a low level resolution, while objects which are close together need to be 'focussed' on with a higher resolution. This is like using smaller and smaller scale maps to find which of two streets (in some town, in some country) is further to the North.

It is certainly worth keeping discrimination trees in mind as a possible component (perhaps in changed form) of more sophisticated models. More generally, I think there are good reasons for supposing that some kind of hierarchical representation for an ordered series (which relates parts of a series together at a more abstract level than individual object pairs) may be useful but need not take the form of a symmetrical binary tree.

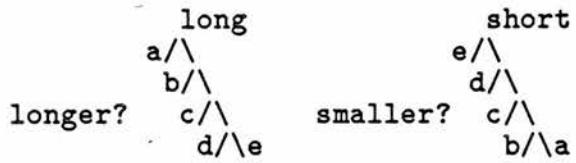


Figure 6–4: Asymmetric discrimination trees

6.4 Linking and Formalising the models

There are a number of connections between these models which become apparent when they are formalised enough to allow computational implementation. One, perhaps counter-intuitive, finding is that there are so many ways of representing the humble series. Some of these can be regarded as computationally isomorphic, but others appear to have differing procedural implications. In any case, at this level of description it ceases to become safe to draw the distinction between operations upon a representation and the representation itself.

It is interesting to note that Breslow's model can be seen as a special case of discrimination trees where the tree is completely lop-sided as shown in figure 6–4. To make an isomorph of the contiguity model, however, it is necessary to have two trees, each constructed in a different direction but representing the same series. One tree is used with each comparative.

Breslow's model can also be captured by a set of logical rules, as pointed out by Bundy[5]. The rules, shown in table 6–5, need a procedural interpretation, as Prolog clauses for example, and then they will produce an asymmetric distance effect such as Breslow would predict (for one comparative). The notation for this and the following models is explained below.

[5]Personal communication

6.4.1 Notation

The notation is Horn clause logic adapted for a standard character set and written in Kowalski form with the head of each clause on the left. This is preferred to standard logical notation as it has a standard procedural interpretation (Kowalski, 1979) which is intended as part of the models. The examples, therefore, run as Prolog programs with minor syntactic changes. See (Clocksin & Mellish, 1981) for an introduction to Prolog.

Constants and functors are written in lower-case and variables in upper-case. Unfortunately, this means that the convention used throughout the rest of the thesis, of writing five-term series items in upper-case italics, needs to be changed here. The series is now written 'a' to 'e'.

Arrows indicate implications, capital letters indicate variables, '&' indicates conjunction and 'v' indicates disjunction. An '=' sign indicates that the structures on either side are unifiable and '\ =' (not equal) indicates that unification fails. A list enclosed in square brackets is 'syntactic sugar' (shorthand) for a nested list structure made up of binary units as shown in table 6–6. A vertical bar within a list indicates that the remainder of the list (also a list) lies to the right of the bar.

In Bundy's representation, the series is explicitly represented as a **list** which is 'scanned' from beginning to end until one of the items to be compared is found. **Lists** are common computational/logical structures used for representing series, sets and bags; and they have an inherently asymmetrical nested structure which is illustrated in figure 6–6. In order to complete the logical implementation of Breslow's model using lists, it is necessary to have a reversed object list for use with the converse comparative, 'smaller'. This is shown in (ii). It is difficult to justify such a program on computational grounds as it seems highly redundant, especially for longer series, when 'smaller' questions can be answered by selecting the item which is 'not-bigger'. Predicate (iii) shows how this can be done. The two alternative definitions of 'smaller' predict different behavioural profiles. In the version (iii) with a reversed list, we have a cross-over effect, with a positive or

```
(i) big_to_small([a,b,c,d,e]).

    bigger(X,Y,Choice) <- big_to_small(List) &
            first(X,Y,List,Choice).

    first(X,Y,[Z|Rest],Z) <- X=Z v Y=Z.
    first(X,Y,[W|Rest],Choice) <-  W\=X & W\=Y &
            first(X,Y,Rest,Choice).

(ii) small_to_big([e,d,c,b,a]).

    smaller(X,Y,Choice) <- small_to_big(List) &
           first(X,Y,List,Choice).

(iii) not_bigger(X,Y,Choice) <- bigger(X,Y,Big) &
           (Choice=X v Choice=Y) & Choice \= Big.
```

Figure 6–5: (i) Bundy's rules for Breslow's model, (ii) and (iii) show alternative implementations of 'choose smaller'.

negative slope being produced depending on which end of the list is scanned from, but no marking effect. With the 'not bigger' version the curves for both 'bigger' and 'smaller' comparisons have the same slope and do not cross. However, comparisons will always be faster for the more primitive 'bigger' predicate, so there is a 'marking' effect.

If this kind of 'scanning' model is correct then one possibility is that both kinds of strategy are used for the marked comparative by different subjects or at different levels of expertise. However for a set of subjects using a uniform strategy we would expect marking and congruity effects to be mutually exclusive. Marschark and Pavio have argued that according to their 'expectancy' account, congruity and marking will be mutually exclusive (Marschark & Paivio, 1981) but Banks et al have disputed their empirical claim (Banks *et al*, 1983).

An alternative computational representation of a series is to use pointers from each object to its successor together with a note of the first item. This is shown in table 6–4. Although the representation of the series is different, the clauses that do the work are virtually identical. The main difference is that an extra variable is needed for the list case in order to store the part of the list structure

```
(i)  [a,b,c,d,e] = .(a, .(b, .(c, .(d, .(e, []))))))

(ii) [Head|Tail] = .(Head, Tail)

(iii)          root
              a/\
               b/\
                c/\
                 d/\
                  e/\[]
```

Lists are nested structures of arbitrary depth. The dot '.' is a functor and '[]' represents (by convention) an empty list. This does not exhaust possible representations.

Figure 6–6: Three different ways of representing a list

which has not been processed, whereas the remainder of the sequence is implicit when pointers are used. This is offset by the fact that the sequence information is stored in separate predicates for the pointer method so that the search space is larger.

Just as with the list representation, the program does not naturally extend to producing a mirror image distance effect for the comparative 'smaller'. A 'smaller' predicate can be produced by having a reversed sequence or doing a transformation, as before. However, logical pointers of the type above are different from the computationally primitive pointer in that they can be used in either direction. That is, a statement that one item is the successor of another can be used equally well to retrieve the predecessor given the successor. As the series is stored above, it is just as easy to work backwards through the pointers starting from the smallest, and a 'smaller' predicate can be written this way without having to represent the sequence twice.

However, not only will this produce no marking effect but it does not fit the evidence that successor links are asymmetrical and that, in general, subjects cannot access a series with equal facility in both directions. As an informal example, try reciting your telephone number or the letters of your name in reverse order.

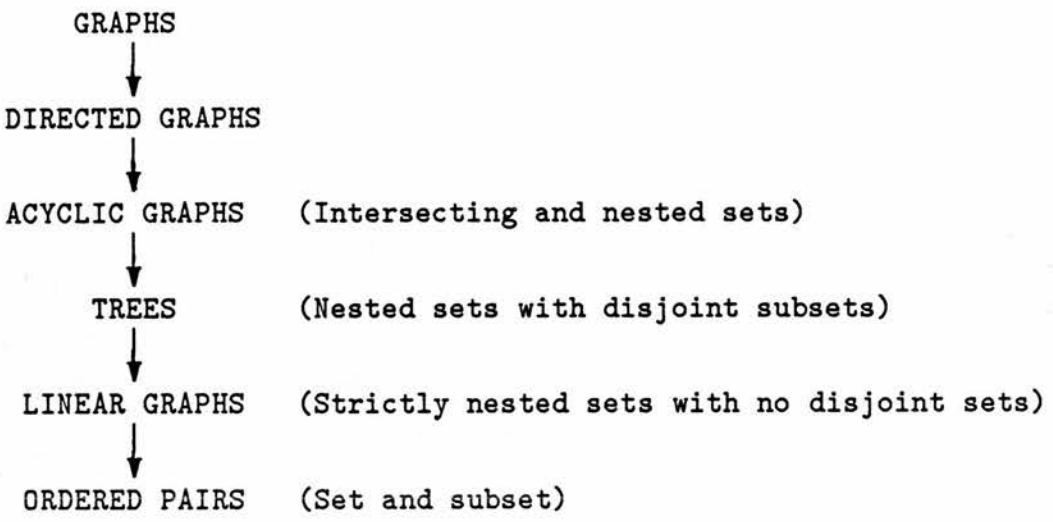
```
biggest(a).

successor(a,b).
successor(b,c).
successor(c,d).
successor(d,e).

bigger(X,Y, Choice) <- biggest(Big) & bigger(X,Y, Big, Choice).

bigger(X,Y, Referent, Referent) <- X=Referent v Y=Referent.
bigger(X,Y, Referent, Choice) <- X\=Referent & Y\= Referent &
        successor(Referent, Next_ref) &
        bigger(X,Y, Next_ref, Choice).
```

Table 6–4: A representation of the Breslow model using pointers

The asymmetry of the primitive pointer arises from a one-to-many relationship. This kind of pointer is like a clue in a treasure hunt that tells the finder where to look for the next clue or for the treasure. An item is paired with a memory location which informs the machine code where the 'pointed to' item is stored. Pointers cannot be traversed backwards without searching through thousands of memory locations. In order to represent a sequence bi-directionally with pointers like these it is necessary create an extra set of pointers. If subjects' successor links are analogous to these then it might help explain directional effects if some or all the pointers in the reverse direction are missing.

Pointers can have further constraints placed upon them than this. An important difference between them and the list structures previously mentioned is that lists are intrinsically constrained to represent linear orders whereas pointers can represent any kind of graph consisting of nodes interconnected by arcs. Viewed like this, linear orders are seen as a special case of graphs. The more general hierarchy of graphs can be depicted as shown in figure 6–7 along with an analogous hierarchy for nested structures.

The case we have been interested in so far is that of the linear graph which is isomorphic with a set of strictly nested sets or a list. A list is a highly constrained structure and to achieve the same effect in a logical model extra axioms have to be added which, effectively, prevent an item having more than one successor



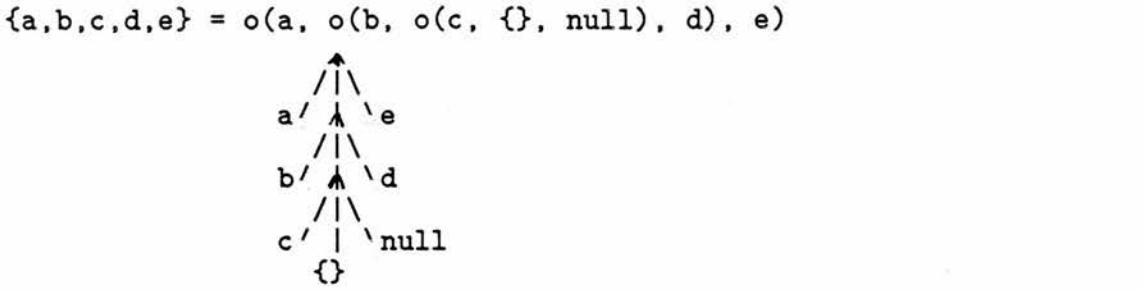
Arrows represent increasing constraint so that items at the bottom can be seen as special cases of the structures above. An analogy between some of the graphs and nested structures is shown in parentheses. Transitive closure is represented by directed paths in graphs and by inclusion in set structures.

Figure 6–7: Linear orders in the context of graphs

or more than one predecessor. In most of the psychological experiments in the transitive inference paradigm subjects were not presented with such a situation which would enable us to judge whether human's successor links are linearly constrained or whether they are equally capable of representing indeterminate or partial orders which have the structure of an acyclic graphs. However, the evidence for the existence of a 'mental line' would seem to suggest that linear representations are a preferred form.

6.4.2 New models

If we allow the pointer of figure 6–4 representation to work in both directions then it is possible to produce a symmetrical distance effect, of the type described by Trabasso, by treating the series as a set of shells as in figure 6–8. This nested structure is much more clumsy than the list, but it can be constructed in an 'ends-inward' fashion which is why it has been included.



This shows how a list might be represented as a nested set of shells so that the middle of the list is the most deeply nested instead of the tail. I have called the hypothetical three place predicate 'o' for 'onion' and the curly brackets are shorthand in the same way as square brackets notate a list. A null atom is needed as well as the empty list.

Figure 6–8: Representing a series as nested shells

The model program in table 6–9 behaves in a similar way to the one in table 6–4 except that the objects are compared with both end items. The program accurately reflects the data as described by Trabasso but it produces no interaction between the position of the item in a series and the form of the comparative (congruity effect). It does produce a marking effect, however as the 'smaller' questions are answered by interpreting the question as a 'bigger-than' one and then transforming the answer. The shell-like structure seems reminiscent of Trabasso's 'end-inwards' principle for constructing series. The program would perform equally well on partially constructed series for which the middle pointers (successor predicates) are missing.

Indirect representation of a series.

The stack model differs in character from the previous accounts in that there is no single preferred 'explicit' representation of the series. Rather, the mechanism provides a framework for indirectly representing the series. It suffices to produce the transitive behavioural profile and yet does not carry with it the implication that the subject should be able to (for example) seriate the objects. The word

```
biggest(a).

successor(a,b).
successor(b,c).
successor(c,d).
successor(d,e).

smallest(e).

bigger(X,Y,Choice) <- biggest(Big) & smallest(Small) &
                bigger(X,Y,Big,Small,Choice).

smaller(X,Y,Choice) <- bigger(X,Y,Bigger) &
                other(X,Y,Bigger,Choice)

bigger(X,Y,Big,Small,Big) <- X=Big & Y=Big.


bigger(X,Y,Big,Small,Choice) <- (X=Small v Y=Small) &
                other(X,Y,Small,Choice).
% If either X or Y is the smallest then select the other.


bigger(X,Y,Big,Small,Choice) <-  X \= Big &
                Y \= Big &
                X \= Small &
                Y \= Small &
                successor(Big,Next_big) &
                successor(Next_small,Small) &
                bigger(X,Y,Next_big,Next_small,Choice).

% If neither X or Y is the biggest nor the smallest then
% find the next biggest and the next smallest and repeat
% the whole procedure (recurse).

other(X,Y,X,Y).  % If X then pick Y
other(X,Y,Y,X).  % If Y then pick X
```

The program instantiates the variable 'Choice' to either X or Y depending on which item is nearest to the biggest object, if the predicate BIGGER is used, or to the smallest item if the predicate SMALLER is used. The program works 'end inwards' on the series; at each stage checking to see if either of the items to be compared is an end item and then chopping the two end items off and repeating the process with the middle section. SMALLER questions are answered by finding which item is the biggest and giving the other item as the choice. Both forms of question give rise to a symmetrical distance effect but SMALLER questions take a fraction longer due to the extra transformation.

Figure 6–9: Ends-inward model of the distance effect

'explicit', above, has been put inside scare quotes because it is used informally, in the way psychologists have tended to use it. Of course the representation *is* explicit when the whole representational system (rules plus control strategy) is considered. One way of thinking about it is a list made from interleaving two others, ordered from different ends of the series (cf Breslow's model).

Of the models which have been analysed, the stack model is closest in kind to the discrimination tree model. In the minimal stack model, each stack can be represented by a tree structure with one level for each rule. When all the rules are of the same form (stacks 1 and 8) the model converges with that of the 'scanning' type models, such as Breslow's. The tree form of stack 3 is represented below. It can be thought of as a special case of discrimination tree where one of the nodes terminates at each branching. The reaction time projections for the binary pairs (previous chapter) are effectively using the depth of the tree as was done for the discrimination tree projections in an earlier section.

```
depth      root
  1         /\e
  2      a/\
  3         /\d
  4      b/\c
```

Stack 3

1) $present(E) \Rightarrow select(E)$
2) $present(A) \Rightarrow avoid(A)$
3) $present(D) \Rightarrow select(D)$
4) $present(C) \Rightarrow select(C)$

Cross-over model

The previous logical models have all had the characteristic that they show either a cross-over effect or marking but never both. For explaining the cross-over effect the models either have the unsatisfactory feature of requiring two independently constructed 'mental lines' or they use logical pointers which do not reflect the overall asymmetry of the series. This model is explained as a derivation of the Breslow rules, although there are other routes to it.

```
objects([a,b,c,d,e]).

bigger(X,Y,Choice) <- objects(List) &
        ranked(X,Y,List,Choice) <(i) then (ii)>.

smaller(X,Y,Choice) <- objects(List) &
        ranked(X,Y,List,Choice) <(ii) then (i)>.

 (i) ranked(X,Y,[Z|Rest],Z) <- X=Z v Y=Z.
(ii) ranked(X,Y,[W|Rest],Choice) <- ranked(X,Y,Rest,Choice)
```

Figure 6–10: Simple rules for cross-over model

Figure 6–10 includes a predicate **ranked** which is exactly the same as the predicate **first** in Bundy's rules for Breslow's model (table 6–5) except that the conditions

```
X\=W & Y\=W
```

have been removed from the body of the recursive call in the second predicate. When this is done then the variable **Z**, and hence **Choice** can logically be instantiated to either **X** or **Y**, the items to choose between. However, according to the procedural interpretation, the *order* in which they instantiate is dependent on the order of the two clauses for **ranked**. If the first choice is accepted as the only answer and the program is procedurally prevented from allowing the other choice to be found (*eg* by using the Prolog cut), then the program will have the same kind of 'reaction time' profile for 'bigger' comparisons as previously, with Bundy's rules. However, 'smaller' comparisons are made differently. This is because the items are not compared as the program recurses *down* the list (tail recursion) but, instead, the work is done on the way up out of the recursion.

An alternative way of viewing this process is that a tree structure (like the one shown in figure 6–6) is searched top-downwards for 'bigger' comparisons and bottom-up for smaller comparisons. A Prolog program which produces this behaviour is included in appendix B.2 and the search trees produced for two 'worst case' comparisons are shown in appendix B.3. These are **d-e** for the comparative 'bigger' and **a-b** for the comparative 'smaller'.

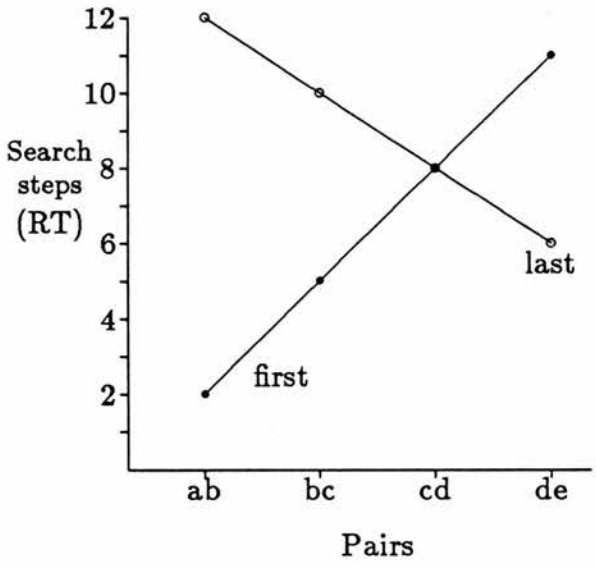


Figure 6–11: Simulated cross-over effect

Figure 6–11 shows how the number of search operations (represented as arrows in the *or* trees in appendix B.3) varies with ordinal position of adjacent pairs. The plot resembles a typical crossover effect with a net time advantage to the unmarked comparative. Another feature of this profile is that the slope of the serial position curve is steeper for the unmarked comparative. By inspection, this seems to be typical of cross-over effects in the symbolic distance literature, for example (Banks & Flora, 1977) but it is unclear whether this is typical of the N-term series task because of lack of evidence. Although this very simple mechanism appears to show both marking and congruity effects, it cannot simultaneously account for the subtleties of the N-term series task. It is concluded that more research needs to be done before the connection between the SDE and N-term series effects can be fully established.

6.5 General discussion

As an account of performance of the N-term series task, the stack model outperforms existing models. A comparison with the binary sampling model, which does not make binary reaction time predictions, was made in the previous chapter. The main strength of the stack model is being able to account for individual variation, but the averaged stack profile also fits the monkey group reaction time profile better than Trabasso's model, Breslow's basic model and Bundy's model. It has been shown that the stack model subsumes the core part of Breslow's model (based on traversal times) and that the assumptions Breslow has to make about the retrieval of end-anchors are unnecessary.

None of the models described in this chapter makes predictions about what would happen in triadic tests without making extra assumptions. In Trabasso's model, all items need to be located in the 'mental line' before a comparison is made, so presumably, triadic comparisons would simply take longer. However, Breslow's model would make equivalent projections to a subset of the stack model comprising stacks 1 and 8, provided that the assumptions of the stack model were also taken aboard. Bundy's discrimination tree model is actually explicitly geared towards making binary choices because of the binary tree representation. Given a triad, some mechanism would have to be provided for dealing with binary subsets. This would simply make the decision process longer, without predicting particular error patterns.

Some alternative representations of series were described above, and formal links between the various models were pointed out. It was shown how a cross-over effect, of the kind produced when question form is varied, can be produced by a very simple representation, in which question form affects search control. This simple model is no good for modelling the details of the five-term series task, however. The connection between the congruity effect found in the SDE literature and the asymmetries found in the five-term series task remains to be established. One possible avenue to explore is that different question forms might

map on to 'winning' or 'losing' in the monkey task. Suppose the instruction 'choose bigger' is analogous to 'winning' (selecting the item which will be rewarded — the tin with a peanut underneath it). Perhaps 'choose smaller' would be analogous to instructing the subject to 'lose' (avoid the 'reward'). There are obvious methodological problems in testing this idea because of the difficulty of motivating the subject to 'lose'. If these problems could be overcome, however, then we might expect a cross-over effect such that 'winning' would be faster towards the rewarded end and 'losing' would be faster towards the unrewarded end of the series. It is not clear whether the stack model could be adapted to deal with such a finding.

Chapter 7

Transitive Inference as Induction

'Proof in logic is only a mechanical expedient to facilitate the recognition of tautology, where it is complicated.' (Wittgenstein 1922 *Tractacus*)

Chapters 5 and 6 demonstrate that the simple stack model has descriptive power with respect to the monkey data. Although evaluation of this kind of performance model could be taken further, this should not be done at the expense of asking why *subjects employ an integrated representation, as opposed to storing adjacent pairs. This chapter argues that performance of the N-term series task is best understood as resulting from an act of Inductive, rather than Deductive, inference. A simple inductive learning algorithm is proposed which builds rule stacks (of the kind previously described), given a sequence of training examples. The chapter appears here, as opposed to in the main modelling section, because it is less empirical and more theoretical and exploratory in nature.*

7.1 Introduction

Although the simple stack model is surprisingly successful in modelling the performance phases of the five-term series tasks, there are likely to be diminishing returns in trying to extend and modify the model to cover more aspects of this data. There would be a danger of over-extending principles expressed in a few simple rules into what is, after all, a complex and dynamic learning situation. What do these rules and principles express? So far, no more than a schema for cheaply processing queries about finite linear orders. It seems unlikely that the stack model can be extended in a principled way without considering wider

issues outside the narrow domain of (deductive) transitive inference. A larger framework is needed within which to explain the following.

Why do subjects make (or appear to make) transitive inferences at all? Perhaps they appear to choose transitively simply because a representation involving a set of avoidance and selection rules is convenient? If this is the case, then why is it convenient, given that subjects presumably *could* represent the training information as discrete pairs? Do the subjects somehow anticipate the presentation of novel pairs and choose a representation accordingly? During testing, the monkey subjects were rewarded indiscriminately, so whatever the reason, it must be entirely to do with representation. Similar problems arise in trying to explain performance on the triadic tests, particularly the spontaneous improvement.

Part of the answer may be that the stack representation is, computationally, cheap enough to be a sensible representation of four pairwise relations. Suppose the four training pairs were totally unrelated to each other. How might the information be represented? Rather than remembering eight items, the most economical solution is to store only four which must be selected or, alternatively, the four which should be avoided. Add to this a mechanism for ordering the comparison of the stored items with test items, and the representation begins to look like the stack model. This still leaves many unanswered questions, however, such as why mixtures of avoidance and selection rules are employed, how the rules become ordered and why the representation should vary between subjects.

I hope to show that the answers to the problems above lie in the fact that the transitivity task is essentially solved by *inductive*, rather than *deductive* processes. We cannot assume either that the subjects have a ready made transitive inference rule or that, even if they had, that they would know to apply it in this context. Without such a rule being added to the axioms representing the task, the transitive relationships between remote pairs simply do not follow deductively from the original pairs.

In order to sort out where and when deductive or inductive inference might occur, we need to consider the task as a whole, from the initial presentations of the 'adjacent' pairs to the improvement in performance on these pairs, through

to the choices made on the test pairs. We need to consider what prior knowledge it would be reasonable to expect the subjects to possess and what individual subjects might learn from the feedback they receive after making every choice. In short (having reached the current level of analysis), the traditional view on the n-term series task, as requiring subjects to make a succession of acts of deductive inference, is no longer helpful. We will employ the "subject's eye" task analysis described at the beginning of chapter 4. The training schema is reproduced below:

A	B	B	C	C	D	D	E
0	+	0	+	0	+	0	+
B	A	C	B	D	C	E	D
+	0	+	0	+	0	+	0

7.1.1 The Space of Possible Inductions

Examples of the different kinds of inductions subjects might make were given in chapter 4, and they are repeated below:

1. $left(C)$ & $right(B) \Rightarrow rewarded(C)$.

2. $present(B)$ & $present(C) \Rightarrow rewarded(C)$.

3. $present(E) \Rightarrow rewarded(E)$

4. $\neg\, rewarded(A)$

5. $present(D)$ & $absent(E) \Rightarrow rewarded(D)$.

6. $\exists x \quad present(x)$ & $\neg\, rewarded(x)$
   (At least one of the choice objects in a trial is not rewarded.)

7. The colours form a series, A, B, C, D, E and from any subset of these colours the item nearest the E end will be rewarded.

8. $present(C)$ & $absent(D) \Rightarrow rewarded(C)$.

How, and at what points, might such inductions be made? Are some inductions likely to be made at the exclusion of others or are the best one found by a process of trial and error? First, we consider the stages at which induction or deduction might occur in the learning process.

Where could deduction/induction occur?

Given that the animal subjects could not be told, even implicitly, that transitive inference was required of them, there are a limited number of possibilities for performing the task:

(a) The subjects possess a transitive inference schema (either innate or learned independently of the task) which they have a tendency to apply to all incoming information.

(b) As above, except that subjects somehow infer that the transitive inference schema is appropriate to this task situation.

(c) The subjects possess some general purpose inductive mechanism (again, innate or learned independently of the task) the application of which results in a transitive choice profile in this task situation.

The first possibility has been independently suggested to me by a number of people to whom I have described this learning problem. It was argued that this would be a good strategy because transitive relationships may somehow be 'common' in nature. Surely, transitive relationships exist only in the representations employed by organisms, not in the world itself. If this is accepted, then the argument must become that transitive representations are frequently usefully imposed on the world. This may be the case, but is a statement lacking explanatory power (it is also somewhat circular). For example, it seems unlikely that subjects could not deal with an analogous task in which they had to learn non-transitive relationships.

The second possibility, that subjects infer the existence of transitive relationships from the task context, appears more plausible. Unfortunately, it is equally lacking in explanatory power without some account of how this might be done. Certainly it does not follow deductively, so subjects would have to induce, after a number of training examples, that a transitive representation can be employed. However, the available empirical evidence appears to show an incremental improvement in performance during acquisition. We do not see a qualitatively different kind of performance suddenly appearing, which is what might be expected if a transitive inference schema were suddenly being brought into play.

Assuming, for the moment, that a transitive schema is employed, how might it be applied?

1. There could be a two stage process in which subjects first induce that there are four pairs to deal with, and then apply a transitive rule to produce some kind of integrated representation. Again, the acquisition profile does not appear to support this.

2. There could be an incremental application of a transitive rule. For example, the rule could be applied when two pairs have been found which it can combine. (Breslow's model is constructed along these lines.)

3. Subjects could inductively infer that there are five objects to represent and proceed to map these onto a (prior) linear representation. The actual mode of linear representation is not relevant here, except to point out that it could even take a form like the stack model's. (Trabasso's account can be viewed along these lines, although he is not clear on when the linear order representation is brought into play.)

None of these possibilities appears particularly plausible for the five-term series task, although the mapping hypothesis does seem appropriate for other forms of transitivity task where subjects have reason to anticipate a linear order, for example, where subjects are asked to make inferences about the relative

heights of a set of people. There is no reason why a single dimension of preference should appear appropriate in this task or what that dimension might be.

If the above line of reasoning is correct, we are left with option (c) above as the only remaining possibility. It is therefore worth exploring the possibility that the subjects must apply some general purpose inductive mechanism(s) to the five-term series task in order to produce a decision procedure for guiding choices. This decision procedure results in a transitive choice profile. Furthermore, the production and the deployment of the decision procedure are not distinct processes. The decision procedure improves incrementally and its success or failure must inform the inductive process.

7.2 Inductive Mechanisms

In the previous section, the induction problem has been characterised as finding the right abstraction in the space of possible inductions and that a great deal depends on the language of abstraction. The philosopher John Stuart Mill described an analogous situation facing any observer of the world:

> 'We must observe that there is a principle implied in the very statement of what Induction is; an assumption with regard to the course of nature and of the order of the universe, namely, that there are such things in nature as parallel cases; that is what happens once will, under a sufficient degree of similarity of circumstances, happen again, and not only again, but as often as the same circumstances occur. This, I say, is an assumption involved in every case of Induction. And, if we consult the actual course of nature, we find that the assumption is warranted. The universe, as far as is known to us, is so constituted that whatever is true in one case is true in all cases of a certain description; the only difficulty is to find what description.' (Gregory, 1981, Quotation from J. S. Mill (1843) *Philosophy of Scientific Method*)

For our purposes, however, the 'description' must include the subject's *actions* as well as the subject's representation of the environment. The purpose of subjects internal representations must be to actively guide behaviour rather than to passively represent the world. The language of abstraction must therefore be

constrained by these requirements and the subject's own representation of the task as discussed previously.

7.2.1 A generic multiple choice task

Let us assume that the generic form of task is to discover a decision procedure for a set of multiple choice situations and that there is no reason to assume *a priori* any particular symbolic relationship between the various choice situations. The target procedure must bottom out at some level that enables the subject to reach out and choose a physical stimulus (object). The first phase of induction must therefore involve discovering a dimension of variation among the stimuli or some way in which they can be categorised. In this particular task there are only two obvious candidates (apart from superfluous features cuch as incidental noises, specks of dirt, or whatever):

1. Objects could be identified by the property *location* from a set of two (left and right).

2. Objects could be identified by the property *colour* from a set of five (A to E).

Note that these properties are not absolute (context free) and that individual properties must generally be defined relative to a superset. This is not so obvious in the case of colours which we tend to think of as ready made categories. However, imagine having to perform the task with a set of non-saturated colours (purple, red-brown *etc*).

Given a means of identifying objects, the choice action can be guided. There are two possible primitive tactics:

1. *Selection* (+*ve* search): If an object can be found with an identifying property which is likely to be rewarded (in the context of the properties of the alternatives), then choose that object.

2. *Avoidance* (*−ve* search): If an object can be found with an identifying property which is *un*likely to be rewarded (in the context of the properties of the alternatives), then choose an object *without* that property.

With binary choices, both tactics are equally effective. Where there are more than two alternatives, the effectiveness will depend on the nature of the task; whether it is more important to avoid bad choices or to identify good ones. To give a naturalistic example; in choosing an item to eat from a number of fruit or berries, *avoidance* criteria might be appropriate for poisonous varieties whereas *selection* (positive) criteria might be appropriate for identifying ripe ones. From this example, it can be seen that it might be advantageous to combine both tactics in a single decision mechanism - negative search tactics for avoiding a minority of easily identified 'bad' choices and positive search tactics for spotting obvious 'good' choices.

However, the current task differs from the more general kind of choice situation described above in that there are only two qualities of choice - rewarded and unrewarded or, from the experimenter's point of view, correct and incorrect. Furthermore, in the triadic tests, *only one item is allowed to be selected during each trial.* This changes the emphasis on the task from a symmetrical one, where *avoidance* and *selection* are equally valid to a situation where the onus is on finding the 'best' item out of three. In this situation there is an advantage in having rules which positively identify correct choices. This point is returned to later when the triadic tests are discussed.

Transitivity is not a necessary property of this kind of mechanism. Even if for every possible (non empty) subset of the superset (all the different fruits or all the different colours in the task) the decision procedure can specify a 'best' choice, this does not necessarily imply a total ordering on the superset — each subset could be treated differently by having complex conditions for the choice tactics. The mechanism could imply opposing directions of ordering for some

subsets[1]. However, if each hypothesis only refers to one object (as in the stack model), a total ordering will arise.

7.2.2 Hypothesis management

How do the reasoning processes of our imaginary subject continue from here? So far (in phase one) the subject has identified distinguishing features of the stimuli from which it must choose, and has two tactics for choice making at its disposal. Phase two must involve hypothesising relationships between the features of choice situations and choice tactics. This requires (a) a language for describing choice situations (trials), (b) a means of generating hypothetical links between trials and choice tactics and (c) a control strategy for applying and testing the rules.

A language for describing/identifying trials

All trials have objects in the same locations so this property alone cannot be used to distinguish trials. Listed below are some possible ways of categorising trials using colour or a combination of colour and location. Trials could be labeled by:

- No label — the same choice tactic is used on all trials (*eg* always choose left).
- The presence or absence of a particular colour.
- A combination of a colour and a particular location.
- A combination of two colours (present or absent).
- A combination of colours and locations.

These possibilities are in increasing order of complexity and I generated them by simply juggling around with the available properties.

---

[1] *eg*: Choose A over B, B over C and C over A.

Evaluating hypotheses

In general, it may be possible to have rules which work for only a subset of choice situations or to have more than one rule which is applicable to choice. Indeed, there may be no decision procedure which gives optimum choices, but any that improves chances of success will be valuable. For these reasons there has to be some means of evaluating the success of individual hypotheses (assigning credit or blame to them) according to some kind of metric of predictive value. This could then be used by a management system in deciding which hypothesis to apply in any given situation. There are a number of standard AI techniques for dealing with this kind of situation (see discussion) and the particular method adopted is not of immediate interest here. However, it should be stated that there must be some way of dealing with situations where no rule is applicable. This could either take the form of a default rule which simply makes an arbitrary (random) choice or a new rule could be hypothesised applicable to that situation. This brings us on to the question of how rules might be hypothesised.

A fairly obvious point to make is that it makes sense to hypothesise and test simple rules before complex ones. This is Occam's razor principle in microcosm. It has been suggested already that rules could be formed by linking trial descriptors with choice tactics. What kinds of rules might this process generate for the five-term series task? Some examples are shown in table 7–1, using '∅' (empty set symbol) to indicate 'don't care' (the absence of a condition).

The last two columns indicate the percentage of (training) trials on which a given rule is applied with correct results and the percentage of trials on which the rule is applicable, respectively. Together, these figures give an indication of the predictive value of the rule. These values could be approximated to empirically by the subject[2]. The first is an indication of the correctness of the rule - as this is a two choice situation any figure less than 50% means that application of

---

[2]There would need to be an additional quantity indicating the size of the sample, and hence the reliability of the approximation

Eg.	*Rule*	*Correct*	*Applicable*
1	$\emptyset \Rightarrow select(Left)$	50%	100%
2	$present(C) \Rightarrow avoid(C)$	50%	50%
3	$present(E) \Rightarrow select(E)$	100%	25%
4	$present(A) \Rightarrow avoid(A)$	100%	25%
5	$present(A) \Rightarrow select(A)$	0%	25%
6	$present(A)$ & $present(B) \Rightarrow select(B)$	100%	25%
7	$absent(E)$ & $present(D) \Rightarrow select(D)$	100%	25%
8	$left(A)$ & $right(B) \Rightarrow select(B)$	100%	12.5%

Figure 7–1: Example hypotheses

the rule is negatively correlated with success. The second value is an indicator of potential usefulness. If a rule was 100% correct and applicable to 100% of trials then no other rules would be needed until novel situations arose. By these criteria, it can be seen that rules 3, 4, 6, and 7 have equal predictive value. However, rules 3 and 4 have simpler preconditions and so, according to this account, are more likely to be discovered first. Rules which have less than 50% success rate (example 5) should be discarded (or the choice tactic converted to its complementary form). Rules with a success rate close to 50% might be useful if their applicability could be appropriately restricted. This could be done by ordering the rules or by adding in extra conditions. This idea of 'uselfulness' is related to the idea of 'minimum entropy' in information theory (Shannon & Weaver, 1964) as used in the inductive rule learning program ID3 (Quinlan, 1979) (see discussion). However, this relies on complex calculations, involving probability theory, over the whole set of examples, and is not really suited to an incrimental learning process of the type described here.

Generation of rules

It is conceivable that rules might be generated by juggling with trial descriptors and various instantiations of the choice tactics. However, this would probably be hopelessly inefficient in all but the most trivial learning situations. This leaves two possibilities, both of them involving the use of features of an actual choice situation for which the existing rules are inadequate. Faced with such a choice, the subject could hypothesise a rule (relating some description of the trial with a choice tactic) and then simply follow that rule. This seems unlikely when the alternative is to simply make an arbitrary choice (randomly, or perhaps using a weak rule) and to construct a hypothesis on the basis of the outcome. The latter leads to faster decisions and generates fewer false hypotheses.

For example, suppose that the subject is faced with a choice between red[3] on the left, blue on the right and green in the middle, and has no existing hypotheses which apply to this situation. If the subject arbitrarily chooses blue on the right, and this turns out to be correct, then it would make sense to hypothesise one (or both) of the following; 'its a good idea to select objects on the right' or, 'its a good idea to select blue objects'. Assume, for the sake of argument, that it is known that the position factor is irrelevant. The rule generated in this case would be:

$$present(Blue) \Rightarrow select(Blue)$$

If, on the other hand, the same arbitrary decision had turned out to be incorrect, the easiest hypothesis to make would be to *avoid* blue objects:

$$present(Blue) \Rightarrow avoid(Blue)$$

A less direct way of capturing the same information would be as a disjunction of hypotheses such as:

---

[3]Colours, as opposed to letters, are used in this example to make it clear that the ordering is unknown.

- $present(Red) \Rightarrow select(Red)$.

- $present(Green) \Rightarrow select(Green)$.

At first sight, it would appear this reason for preferring *avoidance* as a choice tactic only applies when there are more than two objects to choose between. However, there is another reason. In general, an *avoidance* rule is a more direct way of capturing negative information in a rule. To return to the fruits example, imagine facing a choice between two kinds of berries which you have never tried before. On your left, there is a bush full of red berries and on you right, a bush of blue berries. You have no prior experience of berries, other than realising that they are a kind of fruit but, being a hungry bird, you decide to try a red berry. It tastes foul. It would make sense to hypothesise a rule on the basis of some distinguishing feature of the berry (or the bush or whatever), for example, 'avoid red berries'. It would not make sense to hypothesise that blue berries should be preferred in future choice situations.

This example is partly given to help point out the peculiar symmetry of the binary choices in the five-term series (and other psychological tasks). In this paradigm, absence of reward after a selection implies that, had the subject chosen the other stimulus, they *would* have been rewarded. However, we cannot assume that the subject knows this, especially in the early stages of training. In real life, we often face choice situations where (unbeknown to us) either course is bad. It may be misleading therefore, to assume that subjects treat the relationship between 'correct' and 'incorrect' responses in the same way as the complementary relationship between true and false in deductive logic. The example also illustrates how rule *order* could be determined by early subject behaviour. Suppose a subject is beginning training on the fived-term series. If the subject chose A from the set $\{A, B\}$ then the hypothesis that A should be avoided, would be generated. This would very likely end up as the first rule in the stack, as it is always correct. Suppose B were selected (and thus rewarded) however, and the rule $present(B) \Rightarrow select(B)$ were hypothesised. This hypothesis would be

disconfirmed as soon as the pair $\{B, C\}$ arose, and so the final stack order would depend more on responses to other training pairs.

7.2.3 The control strategy

Assuming that hypothesised rules are generated along the lines described above, there remains the problem of applying them and managing their evaluation. I have argued for an empirical evaluation strategy in which rules have an associated 'predictive value', although I have not specified the mechanism for computing such values. Given such values, one control strategy would be to find all the rules which are applicable in the given choice context and then to apply the rule which has the highest value. The value of this rule would then be updated on the basis of the outcome of the decision. Such a mechanism would not have the characteristics of the stack model.

Another possible mechanism would be to approximate to the 'predictive value' system by simply ranking the hypotheses. Although some information about the rules would be lost, this system is computationally much cheaper; rather than finding all the potentially relevant hypotheses, the rules can be searched from the top until a relevant one is found. This, essentially, is what happens in the stack model. During the learning process, incorrect decisions would lead to re-orderings of the hypotheses, or removal of bad rules.

A third possibility is to combine the best features of both systems. Each rule has an associated empirical value but the rules are also ranked for the purposes of efficient, run of the mill decisions. The ordering of the rules would have to be periodically checked to make sure it reflected the rule values. This strategy would combine the cheapness of the stack (in terms of speed and memory load) as a mechanism for making choices with the sensitivity of the empirical value system for learning purposes. There could even be two levels of decision making; an initial, fast solution which could be employed under time pressure and a slower, more accurate decision which could override the first.

These ideas could perhaps be furthered by attempting to build a computational model of induction in multiple choice situations. It seems sensible to start with the simplest possible kind of control strategy (the second one proposed above), employing a ranking of hypotheses. If we are to use the five-term series task as the modelling domain then there is a second reason for preferring this approach - we already have a model of the end-product of the hypothesised inductive process in the form the stack model. Finally, it may well be the case that evidence from a single kind of experiment (however detailed the analysis) would not be enough to support the level of complexity implied by a more general account of induction in choice situations.

7.3 An approximate model of induction in the five-term series task

In order to create a first-order model of the learning processes that take place during the acquisition phase of the five-term series task, a number of simplifying assumptions have been made.

1. The imaginary subject has identified *colour* as the feature to discriminate objects with. Only hypotheses of the form:

   - $present(colour) \Rightarrow select(colour)$.
   - $present(colour) \Rightarrow avoid(colour)$.

   - are considered. In terms of the general model, hypotheses about other features have been eliminated as candidates and hypotheses about combinations of features are not yet considered at this phase. As shorthand for these two types of rule, $select(colour)$ and $avoid(colour)$ are used respectively (the preconditions are the same in each case).

2. The choice tactics are implemented in a particularly simple way, which is a special case for binary choices. The trial is represented by a set of two colours and avoidance of one automatically leads to the choice of the other.

3. Hypotheses are ordered in a strict ranking and this ranking determines the semantics of individual hypotheses. In any given trial, the highest applicable rule in the stack determines the outcome of the decision. In other words, the condition of each hypothesis is the conjunction of its own precondition and the negation of each precondition higher up in the stack.

4. Perfect information is assumed so that if a hypothesis gives rise to a single incorrect decision then it is assumed to be at fault.

Quantifying the learning problem

It can be seen that these assumptions reduce the problem to generating and ordering hypotheses until satisfactory performance is achieved on the training pairs. What magnitude of problem is this in computational terms? One way to think about the problem is to imagine a dumb trial and error mechanism which randomly generated a different rule stack each time the existing stack generated an error. Given five possible colours and two possible hypothesis forms for each colour, there are ten possible hypotheses from which to build a stack. Even assuming that repetition of hypotheses is eliminated within stacks and that each stack contains sufficient rules (four), then this gives us a space of $10 \times 9 \times 8 \times 7 = 5040$ functionally different stacks. As only sixteen of these perform correctly on all the training pairs (chapter 5), it would typically take thousands of trials before performance was obtained.

Clearly, rather than throwing away the whole stack when an error is obtained, the control strategy should be more conservative. The remaining assumptions deal with the generation of hypotheses (5 & 6) and reordering of rules (7 & 8).

5. If no hypothesis is applicable to a given trial then a guess (random choice) is made.

6. New rules are hypothesised after guesses. Following a correct guess of a colour, x, the rule $select(x)$ is hypothesised. Following an incorrect guess, the rule $avoid(x)$ is generated.

7. New hypotheses are added to the bottom of the stack (adding rules higher in the stack would change the meanings of those below).

8. Faulty hypotheses are simply deleted from the stack (alternative methods of reordering the rules involve more book-keeping).

With this simple mechanism, the stack order is only changed by the process of adding new rules at the bottom and removing unsuccessful rules from higher up. This works fine here because it is so easy to generate rules in this simple domain that nothing much is lost if a rule is thrown out. A more sophisticated mechanism could reorder existing rules and only throw out failures if they are already at the bottom of the stack.

7.3.1 An implementation

These simplifying assumptions are sufficient to specify an algorithm for acquiring rule stacks. Such an algorithm has been implemented as a Prolog program (Appendix: C.1). A typical sample run of the program is given in table 7–1. Each row shows the trial presented, the item chosen, the method by which the item was chosen, the feedback given to the subject and the resulting modified stack. The program was started with an empty set of rules and was given training pairs in random order until one of the correct stack forms was produced. After this point, further trials would all be responded to correctly.

In order to build up a picture of the performance characteristics of this algorithm, a program was written to run the algorithm many times over, with randomly selected trials on each run. This program also added up the number of correct answers given by the program (regardless of whether they were guesses or from rules) so that acquisition curves could be plotted. Over 50 runs,

```
?- go.
[c,d], d, random, good, [sel(d)]
|:
[a,b], a, random, -bad, [sel(d),avd(a)]
|:
[d,e], d, by(sel(d)), -bad, [avd(a)]
|:
[d,e], d, random, -bad, [avd(a),avd(d)]
|:
[c,d], c, by(avd(d)), -bad, [avd(a)]
|:
[a,b], b, by(avd(a)), good, [avd(a)]
|:
[d,e], e, random, good, [avd(a),sel(e)]
|:
[b,c], c, random, good, [avd(a),sel(e),sel(c)]
|:
[d,e], e, by(sel(e)), good, [avd(a),sel(e),sel(c)]
|:
[d,e], e, by(sel(e)), good, [avd(a),sel(e),sel(c)]
|:
[d,e], e, by(sel(e)), good, [avd(a),sel(e),sel(c)]
|:
[c,d], c, by(sel(c)), -bad, [avd(a),sel(e)]
|:
[b,c], b, random, -bad, [avd(a),sel(e),avd(b)]
|:
[a,b], b, by(avd(a)), good, [avd(a),sel(e),avd(b)]
|:
[d,e], e, by(sel(e)), good, [avd(a),sel(e),avd(b)]
|:
[c,d], d, random, good, [avd(a),sel(e),avd(b),sel(d)]
|:
[d,e], e, by(sel(e)), good, [avd(a),sel(e),avd(b),sel(d)]
|:
```

This is was produced by the program in C.1, with a training schema analogous to the one given to monkeys. On each line is shown a randomly selected training pair, the item chosen, the method of choice, the feedback obtained and the revised stack. Choices are made on the basis of the stack shown in the row above, starting with an empty stack. The rules are abbreviated to avd(Item), for *present*(*Item*) $\Rightarrow$ *avoid*(*Item*), or sel(Item) for *present*(*Item*) $\Rightarrow$ *select*(*Item*). The stack is displayed horizontally, with the top rule on the left. The feedback corresponds to reward and non-reward. A working stack is produced on the penultimate line, and correct choices are produced thereafter for all pairs (the output has been truncated).

Table 7–1: Sample output from learning algorithm with random trials

the mean number of trials that the program took to build a correct stack was 30, with a standard deviation of 17. This is obviously much better than a 'blind' search as described above. It is also much faster than the actual subjects, but it must be remembered that the problem has been greatly simplified[4].

Figure 7–2 shows the acquisition curves generated as a result of running the program on thousands of trials. Each graph shows the percentage of incorrect choices (out of the total number of choices) for each of the adjacent pairs. The number of times the program was run is equivalent to the number of artificial subjects and the number of trials indicates the point at which the errors were counted for each artificial subject. The four graphs show the 'group' curve after the first 3 trials, after the first 5 trials, after the first 15 trials and the first 30 trials, at which point most subjects reached perfect performance.

It can be seen that the performance immediately takes on the characteristic inverted 'U' shape after only a few trials. The curve starts off almost flat, with about a quarter of the errors falling to each pair, and becomes more pronounced as acquisition progresses. These curves can be compared with those for the monkeys in chapter 3. However, the program does not show a linear serial position effect, as the subjects did on the first block of trials, but this is almost certainly due to the fact that the pairs were not given to the subjects in random order to start with.

7.3.2 Indeterminate training pairs

The program is not restricted to dealing with this particular training set, however. Appendix C.2 shows the stacks generated for training pairs implying a partial order (not implying a total order — the information about the series is indeterminate). Most of the resulting stacks cause the interpreter to behave as

---

[4]Even so, the algorithm is not as efficient as it probably could be because it throws out faulty hypotheses instead of lowering their priority.

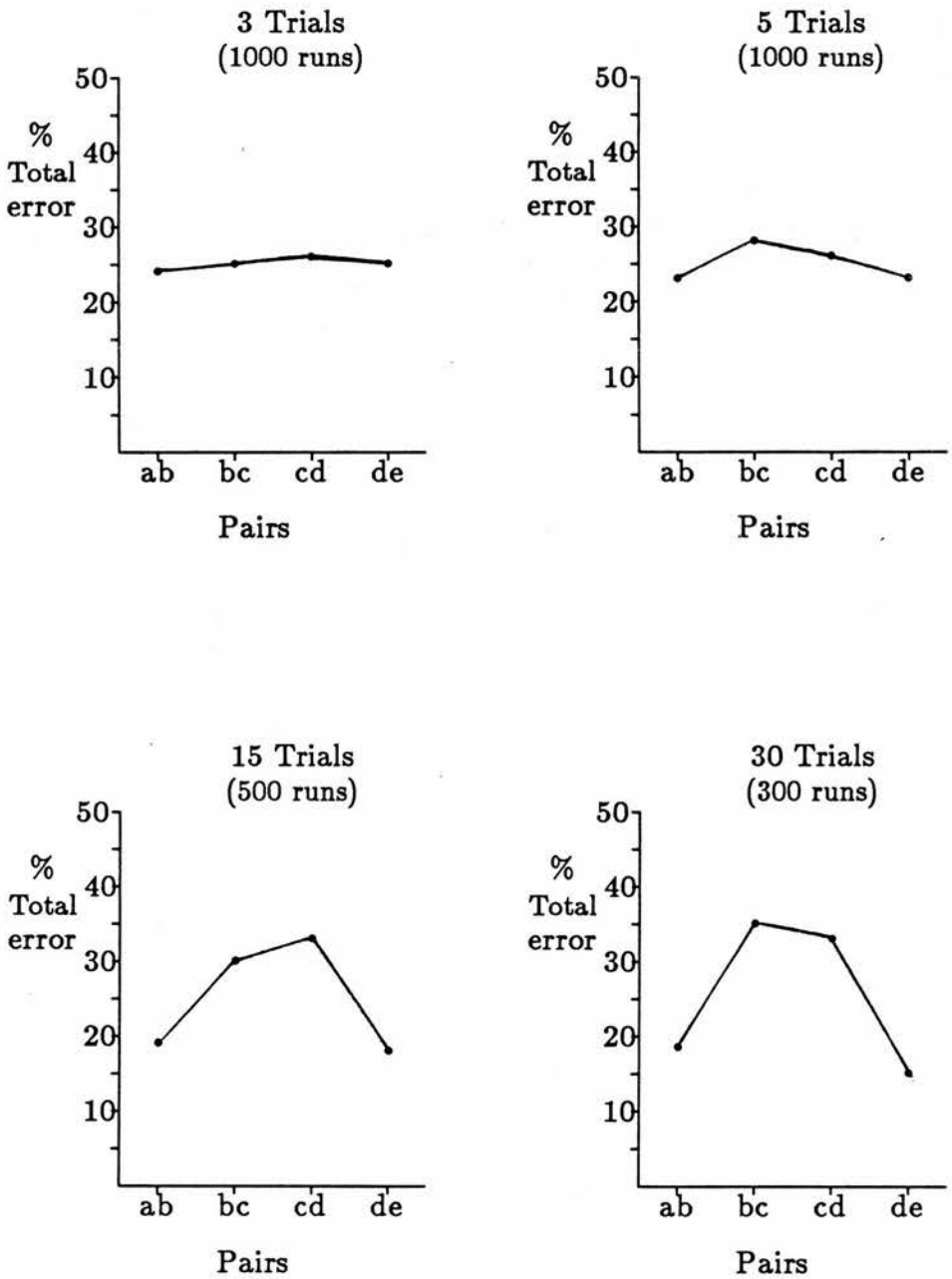
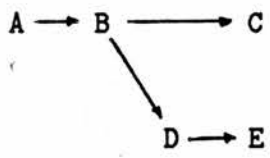


Figure 7–2: Simulated acquisition curves (first N runs).

if the order had been completely linearised (into one of the possible total orderings). However, an unanticipated result was that some of the generated stacks (those with three rules) effectively retain some of the indeterminacy by referring to one less item. The training instances used in the example can be represented by the partial order shown below:



The learning mechanism either collapses this into a total order, for example:

A → B → D → E → C

or it effectively puts two items in an equivalence class, which can be represented as follows:

A → B → (D & C) → E

This retains the ambiguity between C and D but imposes an ordering between C and E which is not inferrable from the training set. I have (informally) observed this kind of partial linearisation in adult's drawings when they are asked to represent indeterminate spatial layouts on the basis of sets of pairwise relations such as 'the cup is to the left of the bottle'. McGonigle and Wright have noted that subjects given indeterminate descriptions will often map the items onto a linear order if retaining the ambiguities is not a task feature (see (McGonigle & Chalmers, 1986)). However, there is currently no evidence available on how monkeys would treat indeterminate training pairs.

7.4 Alternative acquisition mechanisms

The above acquisition algorithm has been presented as a derivation of a philosophical outlook on induction, a kind of 'thought experiment' involving choice tactics in multiple choice situations, and the gross phenomena associated with the acquisition phase of N-term series tasks. However, given the more limited goal of giving an account of the generation of *avoidance* and *selection* rules, there are other possibilities.

7.4.1 The Generalisation heuristic.

This assumes that production rules are the default mode of storing pairwise information and that rules which are initially restricted to dealing with one of the training pairs become 'generalized' to deal with a greater number of situations without leading to incorrect performance on any of the other training pairs. The motivation for this process is hypothesised to be efficiency - four specific (detailed) rules which are each independently correct are effectively converted to a single procedure where each rule contains less detail but correctness is preserved overall by imposing an ordering upon the rules. For example, the rule $present(B) \Rightarrow select(C)$ can be generalised in two ways, by replacing either B or C with a variable x:

- $present(x) \Rightarrow select(C)$
- $present(B) \Rightarrow select(x)$

It is implicit here that the variable is to be matched with a different item to the one referred to in the rule, and this is formalised more rigorously below. However, the main idea is that a rule relating any training pair of items can be generalised by making either its precondition or its post-condition less specific. This is analogous to Generalisation learning in previous production system models. The second and third equation below can be obtained from the first by replacing one

of the items with a variable. The left hand side of the equations acts as a task specification and must be matched against a particular trial and the right hand side specifies the action. Note that these rules are to be regarded as equivalent to the stack model ones.

$$trial(B, C) \rightarrow pick(C) \tag{7.1}$$

$$\forall x : trial(x, C) \rightarrow pick(C) \tag{7.2}$$

$$\forall x : trial(B, x) \rightarrow pick(x) \tag{7.3}$$

Rules about four training pairs therefore lead to a 'parent' set of eight possible rules. These are laid out below using the shorthand 'select(C)' for rules of form 7.2 and 'avoid(B)' for rules of form 7.3.

Original rules	Generated rules	
$trial(D, E) \rightarrow pick(E)$	$select(E)$	$avoid(D)$
$trial(C, D) \rightarrow pick(D)$	$select(D)$	$avoid(C)$
$trial(B, C) \rightarrow pick(C)$	$select(C)$	$avoid(B)$
$trial(A, B) \rightarrow pick(B)$	$select(B)$	$avoid(A)$

The reason that the previous mechanism is preferred to this one is that the reason for performing the generalisations appears to be weak and that subjects would have to achieve a near perfect performance before the characteristic 'inverted U' acquisition curve would appear.

7.4.2 Discussion

A mechanism for acquiring rule stacks has been proposed based on a theory about how organisms deal with multiple choice situations. The stack mechanism, proposed in chapter 4, is seen as a special case of a decision procedure which generates an ordered set of hypotheses about a class of multiple choice situations. The hypotheses are seen as rules when the decision procedure becomes perfected, before that point, they are a source of error.

A language for describing hypotheses was suggested, based on the production rule idea of condition-action pairs. On the right hand side of these hypotheses, there are 'choice tactics'. Two basic tactics, *avoidance* and *selection*, are coupled with object features, giving a space of choice actions. On the left hand side, there is a description which must be just sufficient in complexity to discriminate the choice situations which require different choice actions.

Hypotheses are ordered on the basis of their 'predictive value'. In the simple model of the N-term series task, hypotheses are rejected when they lead to an unrewarded choice and generated when existing hypotheses fail to cover a new situation.

The proposed acquisition algorithm is somewhat difficult to relate to existing AI techniques for rule learning even though, at least superficially, it appears to be tackling an analogous problem. The algorithm seems to differ in a significant way from classification algorithms such as Quinlan's ID3 (Quinlan, 1979) in that these require all the examples in advance rather than working incrementally, and thus they would not allow intermediate levels of performance. The algorithm is not readily comparable with focusing (Bundy *et al*, 1985) because it does not learn unique *concepts*. Instead, what the program does is to come up with one of a number of possible representations/strategies which fit the current task constraints and improve performance. There is no unique 'target' concept to be learned. Another difference is in the use of guessing to generate useful information instead of passively accepting examples and counter-examples.

Although concept learning programs appear to be tackling a somewhat different problem, there are some technical points of comparison. The method of hypothesis ordering does resemble that employed in AMBER, a concept learning program written by (Langley, 1981) which acquires simple English grammar rules when given examples and counter-examples of grammatical sentences. In his system, rules are ordered by having an associated priority number, and where more than one rule is applicable to a given sentence, the priority decides which one is used. A rule's priority gets reduced every time blame for an incorrect decision is attributed to it. In concept learning, this solution is generally unsat-

isfactory (Bundy *et al*, 1985), because the same fault might be detected several times before the priority of the relevant rule drops so low that it is never selected, and this slows learning.

Part of the problem may be that collapsing the value of rules to a single priority dimension is inappropriate in a complex system. Even in the relatively simple choice situations dealt with in this chapter, it was suggested that two factors needed to be taken into account in assessing predictive value (section 7.2.2). There is also the related problem of credit/blame assignment in concept learning programs due to the interaction between rules. These are not problems with the decision procedure described in this chapter, because the hypotheses are highly restricted in form and content. The definition of choice tactics means that credit or blame can always be correctly assigned to a rule when feedback is obtained. The simplicity of the preconditions means that a simple rule ordering is all that is needed.

Chapter 8

Conclusions

This chapter draws together the conclusions from the modelling work and suggests further tests and possible extensions of the model. Finally, the significance of the modelling work is discussed with respect to wider issues in development and learning.

8.1 Goal

Chapter 1 introduced the motivation of this research; to identify and model 'basic cognitive skills'. These are skills which cross age and species boundaries, and which could form the building blocks of more complex skills in biological or synthetic behaving systems.

8.1.1 The Selected Domain

Transitive inference, as found in the N-term series task, appears to be one such basic form of reasoning.

1. It is sub-linguistic and cross-species.

2. It is fast, and closed to introspection.

3. It is abstract and thus not domain limited.

Transitive inference is also interesting from the point of view of synthesis (building AI programs), as it is ubiquitous in formal systems.

8.1.2 Previous Approaches to Transitive Inference

In both the N-term series and the SDE paradigm, described in chapter 2, the kind of data and model generated has tended to depend on whether the researcher has been thinking in terms of analogical representations or in terms of (linguistic) comprehension of comparatives. In particular, the N-term series problem appears to lie at an interesting intersection between two disciplines. On the one hand there is a set of theories and experimental data which conceptualises the act of making comparisons as a kind of symbolic extension to perceptual comparisons. On the other hand we have a body of literature treating comparatives as composite linguistic objects which are matched against similar objects stored in semantic memory. Neither approach gives a satisfactory account of all the phenomena. Further lacking, is an explanation of *how* comparatives are dealt with in terms of *process*, and this is where the main contribution of this thesis lies.

Children and Animals

It appears that the ability of subjects to show a transitive bias on the crucial *BD* pair in the five-term series task is a phenomenon which is robust with respect to species, methodology and stimuli. Furthermore, it seems that 'symbolic' differences between the stimuli determine the representation and that it is not crucial to have the presence of a physical metric such as size. Mcgonigle and Chalmers have explored the relationship between child and monkey data in detail (McGonigle & Chalmers, 1984a) and have found a strong mapping with respect to all the major phenomena, including the acquisition curves, the binary and triadic choice patterns and the ordinal distance effect. All researchers now appear to agree that subjects do not simply store the training information as four separate pairs but that some inferential process takes place during training.

Given this strong concordance, and the fact that the monkey studies are the only ones to have a rich data base for each individual subject, the monkey

data provides the only viable basis for a microanalysis based on computational modelling.

The above research leaves an important question inadequately answered: 'What kind of representation do subjects use to integrate the four pairs?' It is inadequate simply to postulate the *existence* of some kind of linear order device; there are many ways of representing order information, as was discussed in chapter 6.

8.2 Computational Microanalysis

A new model of transitive inference in the N-term series task was developed based on a microanalysis of monkey data. The microanalysis breaks down into approximately two stages, though the development of the model was inevitably partially an iterative process. First, a computational mechanism, the *stack model*, was proposed as a candidate for modelling subjects' decision processes during the N-term series task. The mechanism is simple enough to be plausibly employed by animals and can be justified on the grounds of computational efficiency and parsimony. Second, the data was analysed according to this model.

8.2.1 Evaluation of the stack model

Overall, the stack model is a remarkably simple and effective approximation of monkeys' decision procedures during the five-term series task. The data from seven subjects and a number of experiments has been analysed at several different levels. With the possible exception of one subject (Blue), all the analyses lend some support to the model.

In more detail, the stack model appears to be correct in the following aspects:

1. The distinction between *alpha* and *random* types of triads appears to have a great deal of validity, and supports the notion of subjects employing a

combination of avoidance and selection rules. This is also supported by the clustering of RTs of pairs which would be discriminated by the same rule.

2. There appears to be some mechanism by which rules are ordered, but this may not be a rigid stack as originally supposed. This is supported by the ability of the model to account for much of the variation in choice patterns between triads and, to a lesser extent, the overall linear trend in the RT data for all subjects except Blue. The stack model is able to provide a much better chunking for individuals' RT data than ordinal separation. This suggests that the ordinal distance effect is an 'emergent', rather than a fundamental property of the decision mechanism.

3. The improvement in performance observed across the three triadic testing phases can be partly accounted for by increasing iteration in the control strategy for applying the rules. The residual improvement can potentially be accounted for by the stack becoming more 'rigid' (firmly ordered) with usage, although there is no formal model for this.

There are a number of problems with the model however, some of which are summarised below.

1. The model predicts all the choices to go to the α item in *alpha* type triads, whereas a proportion get diverted to β. It seems likely that this is because the assumption about rule order being fixed is partially wrong.

2. Although for most subjects there exists a linear *trend* between RT and depth of rule, there are significant deviations from linearity for three subjects. It was suggested (chapter 5) that subjects may somehow represent the logic of rule ordering independently of the procedure for rule application. This point is returned to below.

Although it is clear that the proposed 'depth effect' is not a unitary phenomenon (but just a trend in the data), the reaction time analysis also shows

that the ordinal distance effect is similarly artifactual, in that the reaction times do not chunk naturally into the four ordinal separation categories. However, there are strong regularities in the data which are incompletely captured by both models. A reasonable explanation seems to be that the regularity is due to the operation of avoidance and selection rules and that the linear trends are due to to the mechanism which applies these rules.

8.3 The Plausibility of the Stack Representation

The stack model is a computationally cheap solution to the problem of comparing items in a linear order. It is thus plausible that it could be a common biological solution.

As an account of performance on the N-term series task, the stack model outperforms existing models. A comparison with the *binary sampling model*, which does not make binary reaction time predictions, was made in chapter 5. The main strength of the stack model is being able to account for individual variation but, in chapter 6, it was additionally shown that the stack model fits the monkey group reaction time profile better than Trabasso's *spatial discrimination model*, Breslow's (basic) *sequential contiguity model* and Bundy's *discrimination tree model*. It has been shown that the stack model subsumes the core part of Breslow's model (based on traversal times) and that the assumptions Breslow has to make about the retrieval of end-anchors are unnecessary. Furthermore, none of the models described in chapter 6 make predictions about McGonigle and Chalmer's triadic tests, whereas the stack model does.

8.3.1 Information Management

The concept of a stack of *avoidance* and *selection* rules proved to be a useful explanatory construct in accounting for performance on the five-term series

problem. What value might the concept have in explaining reasoning at a more general level? Why was the construct useful? Perhaps the initial answer lies not in the organism but in the epistemology of choice situations, as suggested in chapter 7. There are only two basic ways of narrowing down a choice between a number of objects or courses of action; selecting a subset of the alternatives by virtue of some property (or combination of properties) or rejecting a subset. The avoidance and selection rules employed in the stack model are indeed primitive instantiations of these strategies. More complex strategies, including mixtures of avoidance and selection, can be built from the simple ones by combining them with a simple control strategy.

Despite the fact that the bulk of psychological research involves presenting the subject with choice situations, this simple point about avoidance and selection may have often been missed due to the popularity of the standard procedure of presenting *binary* choices, where it is difficult to tell which strategy is being employed. Even with multiple choice situations, ascertaining the strategy is not trivial, and repeated and multiple measures may be needed. Part of the reason for investigators' reliance on two choice situations may be that they have tended to assume that a subject chooses a stimulus for positive reasons, rather than by avoiding alternatives. If this is the case then a valuable contribution of this research is in demonstrating that *subjects may spontaneously adopt avoidance strategies* in binary choice situations.

In multiple choice situations, avoidance and selection strategies do not have equivalent status; they are not symmetrical with respect to a task and one is not necessarily better than another. The monkey subjects had no way of knowing whether they were being rewarded for the selection of one colour or for avoiding the other, and only when they had to choose one out of three objects (in the triadic tests) did the use of selection strategies show an advantage. One can conceive of giving an equally valid transfer task in which subjects would have to choose *two* out of three items, in which case avoidance strategies would have been more efficient, although this has not been tried yet.

Consider the generality of the architecture of the stack model. Given the

idea that subjects were using a combination of avoidance and selection rules to achieve their performance, we initially tested the simplest form of control that would allow correct performance of the original task (specifically, putting the rules in a stack and trying them one by one until an applicable one is found). This architecture can be generalised as consisting of a set of *hypotheses* about what aspects of the choice situation are good predictors of reward (including non-reward), together with a *control strategy* for hypothesis testing.

This characterisation was used as the basis for a model of the process of learning to perform the N-term series task. This was implemented as a computer program, described in chapter 7. For example, if a blue item is chosen and not rewarded, the algorithm generates a hypothesis that blue should be avoided, which is then provisionally added to the rule set. Only during correct performance can the component hypotheses be regarded as rules. The algorithm acquires the series in an 'end-inwards' fashion, as described in the N-term series literature, and constructs stacks which depend on the previous history of choices, thus potentially accounting for the variation amongst subjects. The model has not been tested in more detail than this, but it does lend some contextual plausibility to the stack model.

8.4 Extensions and Further Tests

8.4.1 The Acquisition Model

As it has been discussed above, the acquisition model is considered first.

It would be interesting to extend this acquisition model to be able to cope with the improvement on the triadic tests. This requires additional sophistication, as it is the control strategy which must change and not just the rules. However, this is not incompatible with the general framework proposed above. On transferring to the triadic tests, the subjects get no feedback on their errors so there would be no reason for them to reject any of their existing rules. The

change must therefore be precipitated by *internal* constraints on the decision process. This could simply be a drive to reduce indeterminacy in the decision process; a default preference for decision by rule rather than by a stochastic process.

Before extending the model in this way, however, it would seem prudent to test it empirically in its current form. It is not easy to do this from the current data base. The problem is that, because of the nature of the experiment, no novel pairs were presented to subjects until they had attained near perfect performance on the adjacent training pairs. If the acquisition model were correct, however, the inclusion of remote pairs at an early stage would have no disruptive effect on the learning pattern; they would vindicate or invalidate existing hypotheses in exactly the same way as for adjacent pairs.

It would be interesting to conduct a similar experiment with 'probe' remote pairs being periodically mixed in with the training pairs. The model predicts limited possibilities for responses on the remote pairs for any given state of learning. For example, if a subject had recently learned the pair BC, but not the pair DE, then it would respond consistently to remote pairs containing these items (AC, BD, BE, and CE) either by avoiding B or by selecting C. If the subject chose C from AC (selecting C) then it would be expected to (incorrectly) choose C from CE and to randomly choose between BD and BE. If, instead, the subject chose D from BD (avoiding B), then a correct response would be expected to BE, with random choices to AC and CE. However, there are obvious difficulties in identifying random choices in a learning situation.

There are other variations on the N-term series task which involve alterations to the subjects' training phase, for example, triads could be introduced at an earlier stage. A more radical change would be to train subjects on an indeterminate series, as described in chapter 7.

Finally, there is the question of whether a different mechanism should be considered in place of the simple control strategy of the stack model. Both the choice and the reaction time data suggest that the idea of a rigidly interpreted stack order is incorrect (although a useful approximation). An alternative was

suggested in chapter 7, which is that rules may in fact be ordered by some kind of measure of 'predictive value' rather than a strict ranking. If this is the case, then there may be something more akin to a competition between the rules (for relevance to a trial) rather than a sequential testing of preconditions . A connectionist type of model might be more appropriate, but it would need to explain why there is a strong linear trend between rule value (depth of rule) and reaction time. This could perhaps be related to the time taken for a connectionist network to reach a stable state.

8.4.2 The SDE phenomena

In chapter 6, some alternative representations of series were described and formal links between the various models were pointed out. It was shown how a cross-over effect, of the kind produced when question form is varied, can be produced by a very simple representation, in which question form affects search control. This simple model is no good for modelling the details of the five-term series task, however. The connection between the congruity effect found in the SDE literature and the asymmetries found in the five-term series task remains to be established. One possible avenue to explore is that different question forms might map on to 'winning' or 'losing' in the monkey task. Suppose the instruction 'choose bigger' is analogous to 'winning' (selecting the item which will be rewarded — the tin with a peanut underneath it). Perhaps 'choose smaller' would be analogous to instructing the subject to 'lose' (avoid the 'reward'). There are obvious methodological problems in testing this idea because of the difficulty of motivating the subject to 'lose'. If these problems could be overcome, however, then we might expect a cross-over effect such that 'winning' would be faster towards the rewarded end and 'losing' would be faster towards the unrewarded end of the series. It is not clear whether the stack model could be adapted to deal with such a finding.

8.4.3 The Stack Model

A number of suggestions for further evaluation of the basic stack model were made in section 5.5.1. These are summarised below:

1. There are reaction times on the triadic tests which could be analysed, although the implications of a semi-iterative control strategy need to be worked out in more detail.

2. Recently, video recordings have been made of the monkeys performing binary and triadic tests, and there is potential for using these to evaluate the model. Possible measures are 'thinking' times (reaction times with the physical reaching component removed) head and eye movements (indicating the focus of attention) and hand movements (these might indicate indecision).

3. There is data from children on the binary and triadic tests which could be analysed, although this is less detailed than the monkey data.

4. Some possible experimental tests were suggested:

   (a) 'Neutral' items (without an ordinal position) could be included in the series and used as probes to test for the use of avoidance and selection strategies. This presents some technical problems, but these could probably be overcome.

   (b) Subjects could be tested with quadruplets as well as pairs and triads (triplets) from the series.

8.5 General Discussion

Investigating basic forms of inference is difficult. It is not initially obvious what mechanisms to propose or how to apply the usual criteria of computational efficiency and parsimony, for the simple reason that these acts of reasoning are so difficult to isolate. The line of attack which we have adopted is to start with a simple computational model and a rich empirical database. The model was initially evaluated against coarse descriptions of the data and, as it withstood statistical tests, evaluated against the fine grain data. At each level of description, the predictions of the model were worked out, and presented in a form compatible with the data. This was possible because of the computational nature of the model.

It is not yet clear what the limitations of computational microanalysis are, but it seems likely that it could be applied to forms of reasoning other than simple linear orders. An experiment with indeterminate (partial) orders has already been suggested. McGonigle and Chalmers have recently carried out an experiment simultaneously involving a symbolic order and a perceptual (size) order on a set of objects, with both monkeys and children. Perhaps an analysis of this might throw some light on the connection between the SDE and N-term series phenomena. Other possibilities stem from the forms of inference mentioned in chapter 2, such as reasoning about equivalences, *eg* (Sidman *et al*, 1982; D'Amato & Salmon, 1985), hierarchies (*eg* class inclusion hierarchies or family trees), sequences (Terrace, 1987) and orders in two or more dimensions, *eg* (Foos, 1980; Mani & Johnson-Laird, 1982).

The motivation for this research was described, in chapter 1, to be 'to work towards discovering the principles of information gain and the control of action in behaving systems'. How does this thesis contribute towards such a goal (apart from developing a methodology)?

Firstly, there is the assumption of the existence of 'primitive' cognitive skills as building blocks for higher order mental operations. It was noted that transitive

inference as observed in children and animals has some of the right features to be a candidate for such a primitive. It is fast (and thus low level), has common (behavioural) properties across species and subjects, and is presumably a useful mechanism 'in the wild' for facilitating decisions involving food preference, social order, *etc.* It was suggested that the same mechanism may be a precursor of more formal logical skills, in particular, formal transitive inference and seriation. This concept of 'primitives' perhaps now needs refining in the light of the modelling attempt. In particular, what is the nature of the 'transitive inference' primitive that has apparently been modelled?

A key problem in understanding the role of the proposed mechanisms is the following paradox. No mention of transitive inference *per se* is made in the task analysis (chapter 4) or in the stack model or learning models. In other words, subjects' strategies can be perfectly well described without explicit reference to transitive inference or, indeed, to any formal logical ability. Is the title of the thesis therefore inappropriate? Is it the case that that the investigation has been of some other form of reasoning?

Yet, at the same time, the monkey's (or child's) behaviour clearly *is* showing (logical) transitive inference from the higher perspective of the *task specification.* That is subjects can show, by their behaviour, that they have inferred relationships between previously unrelated pairs of objects. They are able to choose between abstract symbols *as if* they had performed transitive closure. Do we discount such a perspective on the grounds that subjects are not aware of the experimenter's formal criterion or that the mechanisms 'underneath' employ no specification of transitivity? I think not. More importantly, both perspectives can be resolved if we allow the concept of *pre-logical*[1] inference. This is a category

---

[1]This is distinct from 'para-logical' (deSoto *et al*, 1965) which refers to analogue reasoning devices (spatial or temporal). The term pre-logical is, I believe, consistent with the usage by (McGonigle & Chalmers, 1986), and the ideas described here are an attempt to give their concept a more formal basis. The term 'non-logical' is also fre-

of reasoning which is neither 'logical' nor 'non-logical' (irrational, associationist, stochastic or purely procedural), and has formal properties which are emergent.

Such a category can be justified on developmental grounds. A major philosophical issue in Developmental Psychology is how to explain the emergence of formal (abstract) reasoning skills including, ultimately, mathematical and scientific forms of reasoning. Trying to explain how such development is possible usually raises a similar kind of paradox to the one described above. Either it is assumed that logical skills are innate (and thus available all along) or there is the seeming impossibility of a system trying to pull itself up by its own bootstraps; how can logical relationships be comprehended without logic? This is a bit like the chicken and egg problem.

Piaget suggested that children acquire formal skills by 'internalising their own actions'. That is, they are supposed to abstract logical relationships from the environment via their actions upon it. This explanation is somewhat unsatisfactory to the formalist; what is internalisation — a kind of osmosis? There may be some truth in Piaget's analysis, however, in that the production of behaviour is likely to be an important part of the learning process. Faulty reasoning cannot be corrected unless behaviour is produced. The question is, what mediates pre-logical behaviour and gives the reference points (in the mind of the child) for teaching to be possible. Clearly, what are needed are stepping stones in the developmental path, and these may take the form of pre-logical skills.

But there is still something missing from our characterisation of pre-logical reasoning. It is not sufficient merely to say that the logical properties must be 'emergent' and not formally specified in the components of the strategy. Almost any form of computation can be said to have an emergent logic in that its (input-output) properties can be described as a formal specification. Furthermore, the traditional idea of logical reasoning is purely deductive. It should be emphasised

---

quently employed, *eg* (Breslow, 1981; McGonigle & Chalmers, 1984a) where pre-logical would perhaps be more appropriate.

that I am using 'logic' here in a wider sense to include inductive processes such as conjecture, which are an important part of the five-term series task.

I suggest the following criteria as necessary features of a pre-logical skill.

1. It must be possible to describe a logical relationship between a specification of the task situation and the behaviour typically generated in the task. More formally, the relationship will be a *theory* conjectured with respect to a *logic* (for example, a subset of first order predicate calculus). There are thus two types of inference taking place — *deductive*, where reasoning is either true or false with respect to the theory, and *inductive*, where relationships and theories are conjectured.

2. The logical relationship must be *simple* enough to fit the description 'primitive'. Arbitrarily complex relationships are not allowed.

3. The logical relationship must have *utility*. That is, the relationship must arise in other tasks (presumably, this will be true of most simple relationships).

4. The task specification must be simple enough not to rule out (too many) other cases where the same logical relationship would hold.

For example, the five-term task can be (briefly) described as the presentation of a set of ordered pairs followed by novel (unordered) pairings of the same items. The behaviour is to order the novel pairs (by choosing one and not the other). The relationship between task and behaviour *can* be described by a simple transitive inference rule in conjunction with a formal logic. This type of relationship is useful for reasoning about prefences *etc*, and the task specification is general enough to fit such cases as may arise. The employment of the transitive theory must be considered as a *conjecture* on the part of the subject because the transitive relations do not follow, from logic alone, from the task specification.

Given the above criteria, it can be seen that although the logic of the skill is in the mind of the observer (experimenter), there is always the *potential* of the

subject recognising, or being taught, some of the logical properties directly. For example, all that is needed for the skill to be considerably extended is (a) for subjects to form an abstract representation of the task (for recognising other tasks of the same form) and (b) to retrieve and apply the appropriate strategy when a task is 'recognised'. If the representation of the task is truly minimal (containing no unnecessary features, such as the particular colours in the five-term series task) then the subject could be said to have learned the logical relationship, in some sense. This allows for the possibility of partially formed logical skills where subjects do not always recognise a new task as being isomorphic to a previous one (something which occurs frequently in adults with more complex tasks). So, again, it can be seen that the distinction between 'logical' and 'non-logical' is not so clear cut as it initially appears. There is the possibility of a system appearing to behave increasingly 'logically' without itself possessing a formal system of deductive logic. However, there are some deep philosophical issues at stake here (about the nature of logic), which are beyond the scope of this research to unravel.

Returning to transitive inference, it seems reasonable to describe this, as observed in the five-term series task, as pre-logical primitive. The relationships which the subjects compute can be described as 'transitive' from the outside, yet it seems unlikely that have a complete abstract representation of the task or access (even tacitly) to a formal theory of transitivity. However, they do transfer from the binary to the triadic tests and so must have a partially abstract task representation. Perhaps the term *proto-transitive* inference would be more appropriate for this kind of behaviour (in which case the title of the thesis would, after all be inappropriate).

If the analysis in the learning chapter (7) is correct, proto-transitive inference is the product of a more general purpose mechanism for making choices in certain kinds of multiple choice situation. The mechanism decides what different choices have in common and hypothesises which features are important. In some particular circumstances, perhaps when the range of distinguishing features is small (colour), this general mechanism produces proto-transitive reasoning by effec-

tively imposing a linear order on the alternatives. The five-term series task can be regarded as artificial in that only adjacent terms in the series are presented during learning. However, there is a 'natural' equivalent in that information about choice situations is unlikely to be complete. The 'task' the behaving system faces is to come up with a useful decision strategy given a subset of the types of choices which may arise.

We have the possibility, therefore, of explaining not only the emergence of a formal reasoning capability, transitive inference, from a precursor 'proto-transitive inference', but also the origin of the latter as a special case of reasoning about multiple choices. Although much research would have to be carried out to evaluate such a hypothesis, the potential rewards are large; if it can be shown how transitive inference develops, then this itself may form a model for understanding how we, and machines, may learn to perform more complex forms of reasoning.

References

Banks, W.P. and Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, 3(2):278–290.

Banks, W.P., White, H., Sturgill, W. and Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgements. *Journal of Experimental Psychology*, 9(4):560–582.

Breslow, L. (1981). Re-evaluation of the literature on the development of transitive inference. *Psychological Bulletin*, 89:325–351.

Bryant, P.E. and Trabasso, T.R. (1971). Transitive inferences and memory in young children. *Nature*, 232:456–458.

Bryant, P.E. (1974). *Perception and Understanding in Young Children.* Methuen & Co, London.

Bundy, A. (1986). What kind of field is artificial intelligence? In Bundy, *Proceedings of the Workshop on the Foundations of Artificial Intelligence.*, New Mexico, February 1986, To appear.

Bundy, A., Silver, B. and Plummer, D. (1985). An analytical comparison of some rule-learning programs. *Artificial Intelligence*, 27:137–181.

Clark, H.H. (1969). Influence of language on solving three term series problem. *Journal of Experimental Psychology*, 82(2):205–215.

Clocksin, W.F. and Mellish, C.S. (1981). *Programming in Prolog.* Springer Verlag, Berlin, Heidelberg, New York.

D'Amato, M.R. and Salmon, D.P. (1985). Cognitive processes in cebus monkeys. In Roitblat, H., T., Bever and Terrace, H.S., (eds.), *Animal Cognition.*, chapter 1, Laurence Erlbaum, Hillsdale, New Jersey.

de Soto, C.B, London, M. and Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2(4):513–521.

Donaldson, M. and Wales. (1970). On the acquisition of some relational terms. In Hayes, J.R., (ed.), *Cognition and the Development of Language*, Wiley, New York.

Foos, P.W. (1980). Constructing cognitive maps from sentences. *Journal of Experimental Psychology: Human Learning and Memory*, 6(1):25–38.

Gregory, R.L. (1981). *Mind in Science.* Vail-Ballou Press, Inc., Binghampton, New York.

Hagert, G. (1984). Modeling mental models: experiments in cognitive modeling of spatial reasoning. In O'Shea, T, (ed.), *ECAI: Proceedings of the sixth European conference on Artificial Intelligence.*, pages 389–398.

Harris, M.R. (1985). *Human processing of information involving ordinal relations.* DAI Discussion paper No. 4, University of Edinburgh.

Holyoak, K.J. and Walker, J.H. (1976). Subjective magnitude information in semantic orderings. *Journal of Verbal Learning and Verbal Behaviour*, 15:287–299.

Hunter, I.M.L. (1957). The solving of three term series problems. *British Journal of Psychology*, 48:286–298.

Huttenlocher, J. (1968). Constructing spatial images: a strategy in reasoning. *Psychological Review*, 75:550–560.

Inhelder, B. and Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* Routledge and Kegan Paul, London.

Johnson-Laird, P.N. (1972). Three term series. *Cognition*, 1(1):57–82.

Johnson-Laird, P.N. (1983). *Mental Models.* Cambridge University Press, Cambridge.

Kosslyn, S.M. (1981). The medium and the message in mental imagery: a theory. *Psychological Review*, 88(1):46–66.

Kosslyn, S.M., Murphy, G.L, Bemesderfer, M.E. and Feinstein, K.J. (1977). Category and continuum in mental comparisons. *Journal of Experimental Psychology*, 106(4):341–375.

Kowalski, R.A. (1979). *Logic for problem solving.* North Holland.

Langley, P. (June 1981). *Knowledge acquisition through error recovery.* CIP Working Paper 432, Carnegie-Mellon University, Pittsburgh USA.

Mani, K. and Johnson-Laird, P.N. (1982). The mental representation of spatial descriptions. *Memory and Cognition*, 10:181–187.

Marschark, M. and Paivio, P. (1981). Congruity and the perceptual comparison task. *Journal of Experimental Psychology: Human Perception and performance*, 7(2):290–308.

McGonigle, B.O. and Chalmers, M. (1977). Are monkeys logical? *Nature*, (267):694–696.

McGonigle, B.O. and Chalmers, M. (1984a). Are children any more logical than monkeys on the five term series problem? *Journal of Experimental Child Psychology*, (37):355–377.

McGonigle, B.O. and Chalmers, M. (1984b). The selective impact of question form and input mode on the symbolic distance effect in children. *Journal of Experimental Child Psychology*, (37):524–554.

McGonigle, B.O. and Chalmers, M. (1986). Representations and strategies during inference. In Myers, Brown and McGonigle, (eds.), *Reasoning and Discourse*, chapter 5, Academic Press, New York and London.

Moyer, R.S. (1973). Comparing objects in memory: evidence suggesting an internal psychophysics. *Perception and Psychopsysics*, 13:180–184.

Newell, A. (1973). Production systems: models of control structures. In Chase, W.G., (ed.), *Visual information processing*, pages 463–526, Academic Press, New York.

Newell, A. (1977). On the analysis of human problem solving. In Johnson-Laird, P.N. and Wason, P.C., (eds.), *Thinking: Readings in cognitive science*, chapter 3, Cambridge University Press, Cambridge, UK.

Pavio, A. (1975). Perceptual comparisons through the mind's eye. *Memory and Cognition*, 3:635–647.

Piaget, J and Inhelder, B. (1969). *The psychology of the child.* Basic Books, New York.

Polich, J.M. and Potts, G.R. (1977). Retrieval strategies for linearly ordered information. *Journal of Experimental Psychology: Human Learning and Memory*, 3(1):10–17.

Potts, G.R. (1972). Information strategies used in the encoding of linear orderings. *Journal of Verbal Learning and Verbal Behaviour*, 11:727–740.

Potts, G.R. (1974). Storing and retrieving information about ordered relationships. *Journal of Verbal Learning and Verbal Behaviour*, 103:431–439.

Pylyshyn, Z.W. (1981). The imagery debate: analogue media versus tacit knowledge. *Psychological Review*, 88(1):16–45.

Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In Michie, D., (ed.), *Expert systems in the micro-electronic age*, pages 168–201, Edinburgh University Press, Edinburgh, UK.

Riley, C.A. and Trabasso, T. (1974). Comparatives, logical structures and encoding in a transitive inference task. *Journal of Experimental Child Psychology*, 17:187–203.

Shannon, E C and Weaver, W. (1964). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Shipley, W, Coffin, J. and Hadsell, K. (1945). Reaction time in judgements of colour preferences. *Journal of Experimental Psychology*, 35:206–215.

Sholz, K.W. and Potts, G.R. (1974). Cognitive processing of linear orderings. *Journal of Experimental Psychology*, 102(2):323–326.

Sidman, M., Rauzin, R., Lazar, R., Cunningham, S., Tailby, W. and Carrigen, P. (1982). A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children. *Journal of the Experimental Analysis of Behavior*, 37:23–44.

Smedslund, J. (1966). The development of concrete transitivity of length in children. *Scandinavian journal of Psychology*, 7:81–92.

Terrace, H.S. (1987). Chunking by a pigeon in a serial learning task. *Nature*, 325(January):149–151.

Trabasso, T. and Riley, C.A. (1975). On the construction and use of representations involving linear order. In R.L., Solso, (ed.), *Information processing and Cognition. The Loyola Symposium.*, Laurence Erlbaum, Hillsdale, New Jersey.

Trabasso, T., Riley, C.A. and Wilson, E.G. (1975). The representation of linear order and spatial strategies in reasoning. In Falmagne, R., (ed.), *Reasoning representation and process.*, Laurence Erlbaum, Hillsdale, New Jersey.

Woocher, F.D., Glass, A.L. and Hollyoak, K.J. (1978). Position discriminability in linear orderings. *Memory and Cognition*, 6(2):165–173.

Young, R.M. (1976). *Seriation by children: An artificial intelligence analysis of a Piagetian task.* Birkhauser, Basel.

Appendix A

Stack Projections

Triad	1		2		3		4		5		6		7		8		*Mean%*	
$\gamma\ \beta\ \alpha$	β	α	β	α	β	α	β	α	β	α	β	α	β	α	β	α	β	α
ABC	0	10	5	5	5	5	5	5	5	5	5	5	5	5	5	5	44.0	56.0
BCD	0	10	0	10	0	10	5	5	0	10	5	5	5	5	5	5	25.0	75.0
BDE	0	10	0	10	0	10	0	10	0	10	0	10	5	5	5	5	12.5	87.5
CDE	0	10	0	10	0	10	0	10	0	10	0	10	0	10	5	5	6.0	94.0
BCE	0	10	0	10	0	10	0	10	0	10	0	10	5	5	5	5	12.5	87.5
ABD	0	10	0	10	0	5	5	5	5	5	5	5	5	5	5	5	37.5	62.5
ACD	0	10	0	10	0	5	5	5	5	5	5	5	5	5	5	5	37.5	62.5
ADE	0	10	0	10	0	10	0	10	5	5	5	5	5	5	5	5	25.0	75.0
ABE	0	10	0	10	0	10	0	10	5	5	5	5	5	5	5	5	25.0	75.0
ACE	0	10	0	10	0	10	0	10	5	5	5	5	5	5	5	5	25.0	75.0

Projected choices from the eight stacks described in chapter 4 with the *naive* interpreter, with 10 trials per triad and assuming 50/50% split for *random* triads. The summaries are used in section 5.2.2

Appendix B

B.1 Discrimination tree program

This program was referred to in section 6.3:

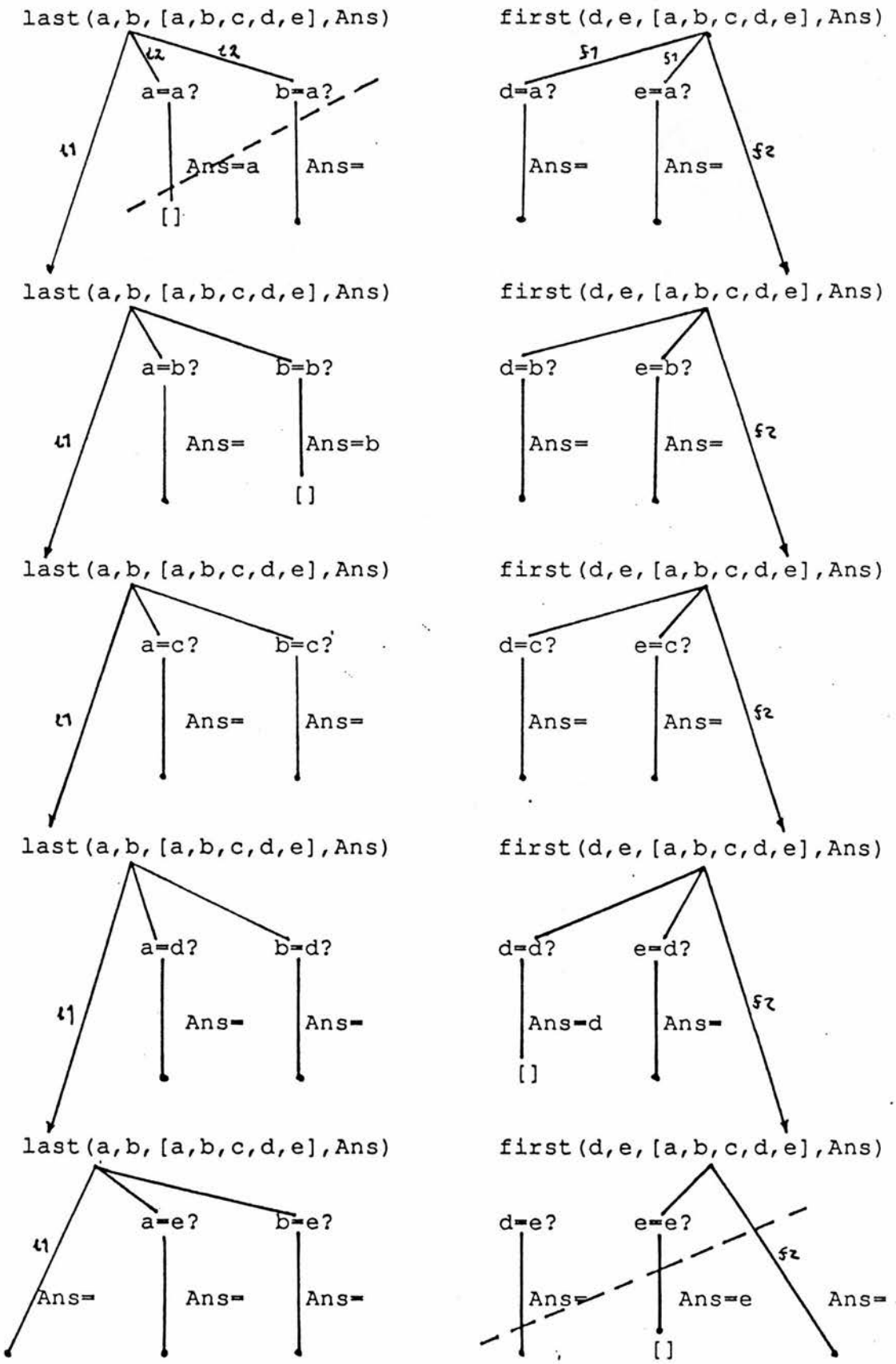
```
Discrimination Net Model of Brendan McGonigle's bigger-than data

Alan Bundy 6.4.82 */

/* top level procedures */

bigger(A, B, Ans) :- /* to find bigger of A and B */
most(big, small, A, B, Ans). /* find which is closer to big pole */


/* General discrimination net */

most(Pole1, Pole2, A, B, Ans) :- /* to find whether A or B is closer to Pole2
*/
category(Pole1, Pole2, A, CatA), /* look up paths of each in tree */
category(Pole1, Pole2, B, CatB),
most(Pole1, Pole2, A, CatA, B, CatB, Ans). /* and compare these paths */

most(Pole1, Pole2, A, [Pole1 | Ra], B, [Pole2 | Rb], A).
/* return A if its path diverges to Pole1 */

most(Pole1, Pole2, A, [Pole2 | Ra], B, [Pole1 | Rb], B).
/* return B if its path diverges to Pole1 */

most(Pole1, Pole2, A, [Pole | Ra], B, [Pole | Rb], Ans) :- /* if paths start
the same */
most(Pole1, Pole2, A, Ra, B, Rb, Ans). /* then recurse on tails */


/* data

Data file for use with mcgon2
Alan Bundy 30.6.82 */

category(big, small, a, [big,big,big]).
category(big, small, b, [big,big,small]).
category(big, small, c, [big,small]).
category(big, small, d, [small,big]).
category(big, small, e, [small,small]).
```

B.2 Cross-over model implementation

```
objects([a,b,c,d,e]).

bigger(X,Y,Ans):- objects(List), first(X,Y,List,Ans).

smaller(X,Y,Ans):- objects(List), last(X,Y,List,Ans).

first(X,Y,[Z|Rest],Z):- (X=Z ; Y=Z), !.      % f1
first(X,Y,[W|Rest],Z):- first(X,Y,Rest,Z).   % f2

last(X,Y,[W|Rest],Z):- last(X,Y,Rest,Z), !.  % l1
last(X,Y,[Z|Rest],Z):- X=Z ; Y=Z.            % l2


% Alternative predicate for SMALLER using a transformation strategy.

smaller(X,Y,Ans):- bigger(X,Y,Big_one),
(X=Ans ; Y=Ans),
Ans\=Big_one.
```

B.3 Or-tree

Search space for first and last in cross-over model:



best case 5
worst case 12

best case 1
worst case 11

Appendix C

Induction

C.1 A program for inferring rule sets from training examples

This the implementation of the program described in 7.3. Adjacent pairs are randomly selected from the series [a,b,c,d,e] and a stack of hypothesised rules are built up. The program does not terminate, but at some point all the generated responses become correct.

```
% Program for modelling the way monkeys learn the five-term series.

% Runs on Quintus Prolog version 2.0. No special features are used
% other than the library program for making random selections.
:- ensure_loaded(library('random.pl')).

go:- loop([],_).  % Starts loop with an empty stack.

% loop/2 makes the program cycle through random selection
% of a pair, making a response and printing output.

loop(Stack,NewStack):- get_pair(Pair),
        respond(Pair,Choice,Method,Feedback,Stack,NewStack), !,
        write(Pair), write(','), tab(1),
        write(Choice), write(','), tab(1),
        write(Method), write(','), tab(1),
        write(Feedback), write(','), tab(1),
        write(NewStack), nl,
        get0(10),
        loop(NewStack,_).


% respond/6 is called with Pair instantiated to a two element
% list and Stack instantiated to a list of rules.

respond(Pair,Choice,Method,Feedback,Stack,NewStack):-
        apply_stack(Pair,Stack,Choice,Method),
        result(Pair,Choice,Feedback),
        revise_stack(Method,Feedback,Choice,Stack,NewStack).
```

```
% apply_stack/4 acts as interpreter for the stack of rules,
% instantiating the third argument to one of the items in Pair
% and the fourth argument to the decision method. The rules are
% abbreviated to avd(Item) (if Item is present then avoid it)
% or sel(Item) (If Item is present then select it).

apply_stack(Pair,[],Choice,random):- !,
        random_member(Choice,Pair).
apply_stack(Pair,[sel(Item)|_],Item,by(sel(Item))):-
        member(Item,Pair),
        !.
apply_stack(Pair,[avd(Item)|_],Choice,by(avd(Item))):-
        remove(Item,Pair,[Choice]),
        !.
apply_stack(Pair,[_|Rest],Choice,Method):-
        apply_stack(Pair,Rest,Choice,Method).


% revise_stack/5 Modifies the Stack on the basis of the outcome of the
% preceding choice and the method employed to make it. If a rule was
% used correctly, no change is made; if a rule was used incorrectly,
% it is removed; if a guess was made a new rule is hypothesised and
% appended to the bottom of the stack.

revise_stack(by(_),good,_,Stack,Stack).
revise_stack(by(Tactic),-bad,_,Stack,NewStack):-
        remove(Tactic,Stack,NewStack).
revise_stack(random,Feedback,Choice,Stack,NewStack):-
        hypothesise(Choice,Feedback,H),
        append(Stack,[H],NewStack).

hypothesise(Choice,good,sel(Choice)).    % Rules hypothesised after
hypothesise(Choice,-bad,avd(Choice)).    % a random choice.


% Training Schema for Monkey Experiment

% For simplicity, items within the pairs always appear in the same
% left-right order, although the program above does not assume this.

get_pair(Pair):-
        random_member(Pair,[[a,b],[b,c],[c,d],[d,e]]).
% Randomly selects a training pair.

% Feedback scheme

result([a,b],b,good).
result([a,b],a,-bad).
result([b,c],c,good).    % There are only two categories of feedback:
result([b,c],b,-bad).    % 'good' is where the response is rewarded;
result([c,d],d,good).    % '-bad' is where there is no reward. Other
result([c,d],c,-bad).    % training shemes are possible.
result([d,e],e,good).
result([d,e],d,-bad).
   % The minus sign is just to make '-bad' 4 characters long
```

```
   % which makes the printout easier to read.

% Utilities

% random_member is a library predicate which
% selects an element from a list pseudo-randomly.

remove(E,[E|T],T):- !.
remove(E,[H|T],[H|L]):- remove(E,T,L).

append([],L,L).
append([H|T],L,[H|R]):- append(T,L,R).

member(E,[E|_]).
member(E,[_|T]):- member(E,T).
```

C.2 Stacks generated from indeterminate training pairs

The following stacks were generated by the above program with a different training schema in which the training pairs were not all adjacent in a series:

```
The training pairs:

if [a,b]  rewarded b
if [b,c]  rewarded c
if [b,d]* rewarded d
if [d,e]  rewarded e

[avd(a),avd(b),sel(c),avd(d)]
[avd(a),avd(b),sel(c),sel(e)]
[avd(a),avd(b),avd(c),avd(d)]
[avd(a),avd(b),avd(c),sel(e)]
[avd(a),avd(b),avd(d)]
[avd(a),avd(b),sel(e)]
[avd(a),sel(c),avd(b),avd(d)]
[avd(a),sel(c),avd(b),sel(e)]
[avd(a),sel(c),sel(e),avd(b)]
[avd(a),sel(c),sel(e),sel(d)]
[avd(a),sel(e),avd(b)]
[avd(a),sel(e),sel(c),avd(b)]
[avd(a),sel(e),sel(c),sel(d)]
[avd(a),sel(e),sel(d),avd(b)]
[avd(a),sel(e),sel(d),sel(c)]
[sel(c),avd(a),avd(b),avd(d)]
[sel(c),avd(a),avd(b),sel(e)]
[sel(c),avd(a),sel(e),avd(b)]
[sel(c),avd(a),sel(e),sel(d)]
[sel(c),sel(e),avd(a),avd(b)]
[sel(c),sel(e),avd(a),sel(d)]
[sel(c),sel(e),sel(d),avd(a)]
[sel(c),sel(e),sel(d),sel(b)]
[sel(e),avd(a),avd(b)]
[sel(e),avd(a),sel(c),avd(b)]
[sel(e),avd(a),sel(c),sel(d)]
[sel(e),avd(a),sel(d),avd(b)]
[sel(e),avd(a),sel(d),sel(c)]
[sel(e),sel(c),avd(a),avd(b)]
[sel(e),sel(c),avd(a),sel(d)]
[sel(e),sel(c),sel(d),avd(a)]
[sel(e),sel(c),sel(d),sel(b)]
[sel(e),sel(d),avd(a),avd(b)]
[sel(e),sel(d),avd(a),sel(c)]
[sel(e),sel(d),sel(c),avd(a)]
[sel(e),sel(d),sel(c),sel(b)]
```